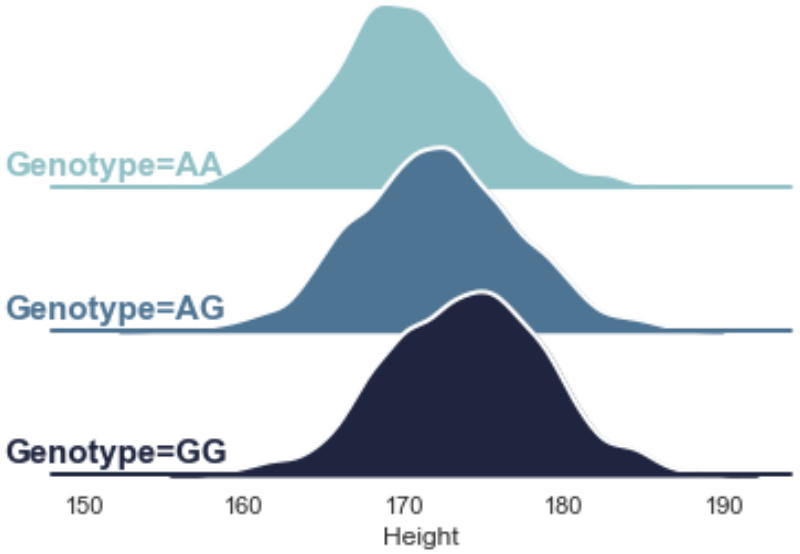
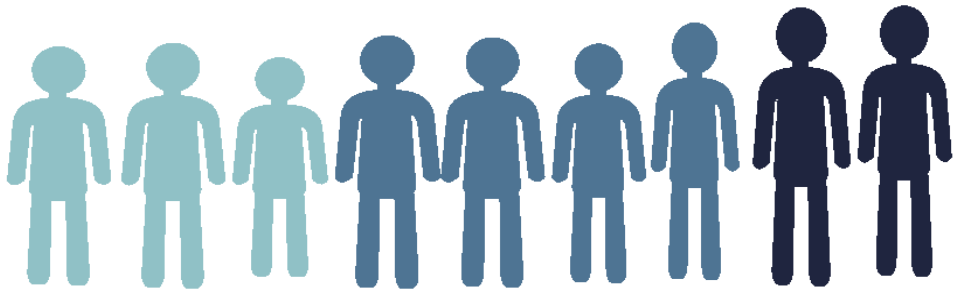
Polygenic Prediction

Aysu Okbay

Vrije Universiteit Amsterdam

a.okbay@vu.nl





Outline

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

Outline

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

Polygenic score (PGS)

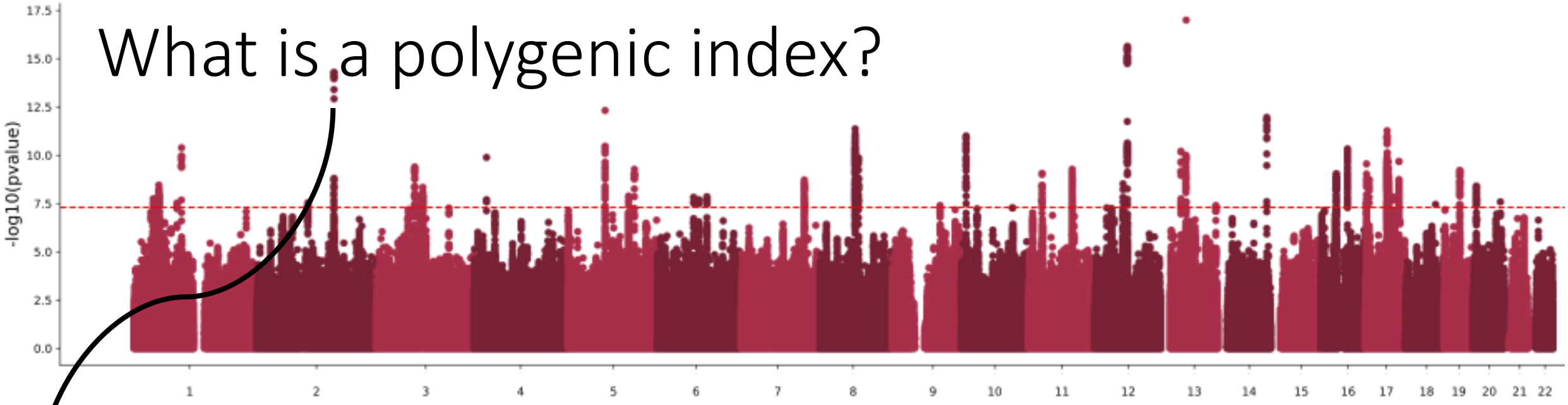
Polygenic index (PGI)

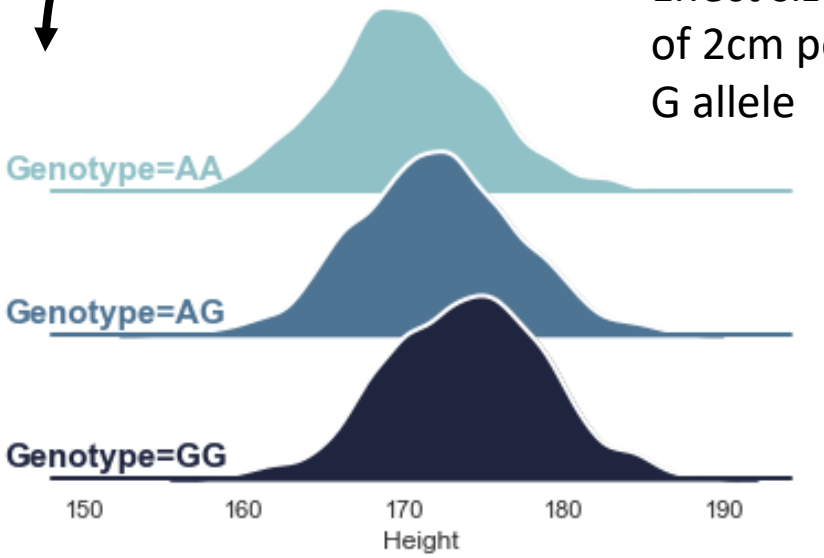
Polygenic risk score (PRS)

Genome-wide score (GWS)

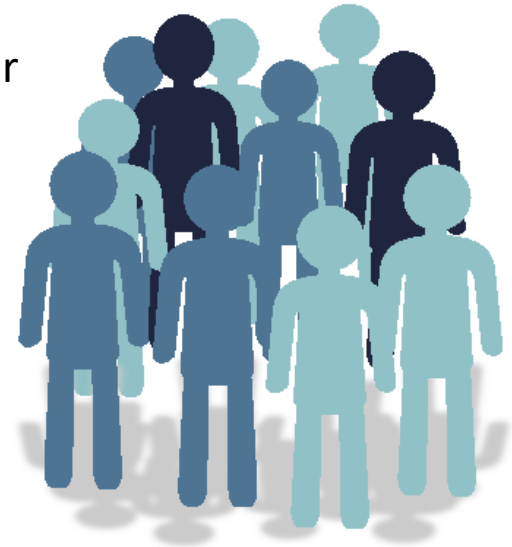
Genetic risk score (GRS)

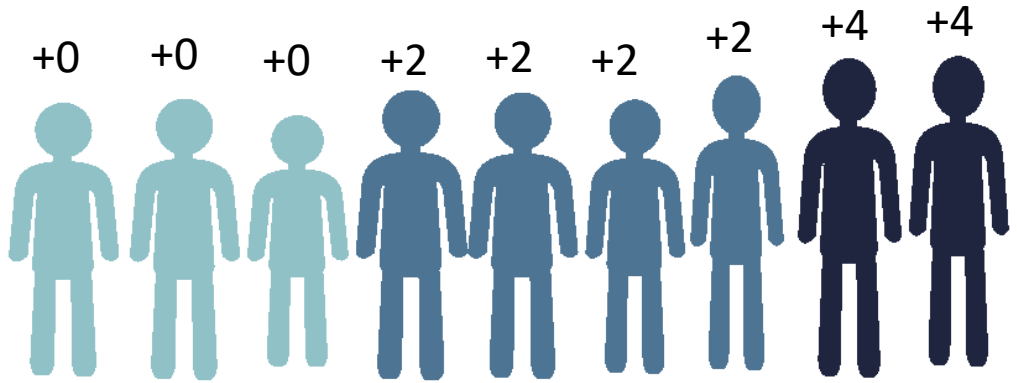
What is a polygenic index?

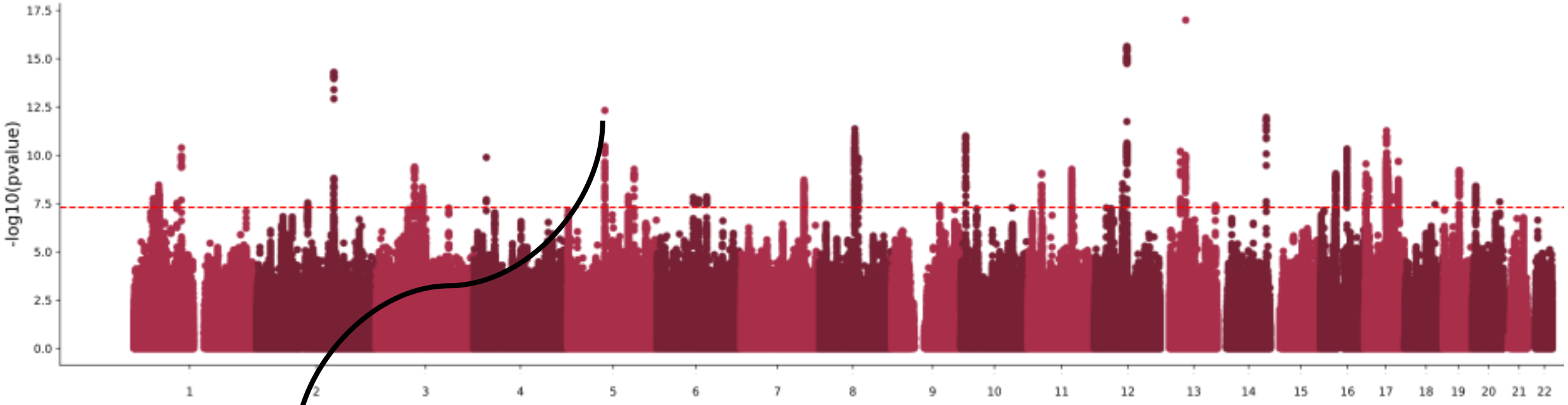




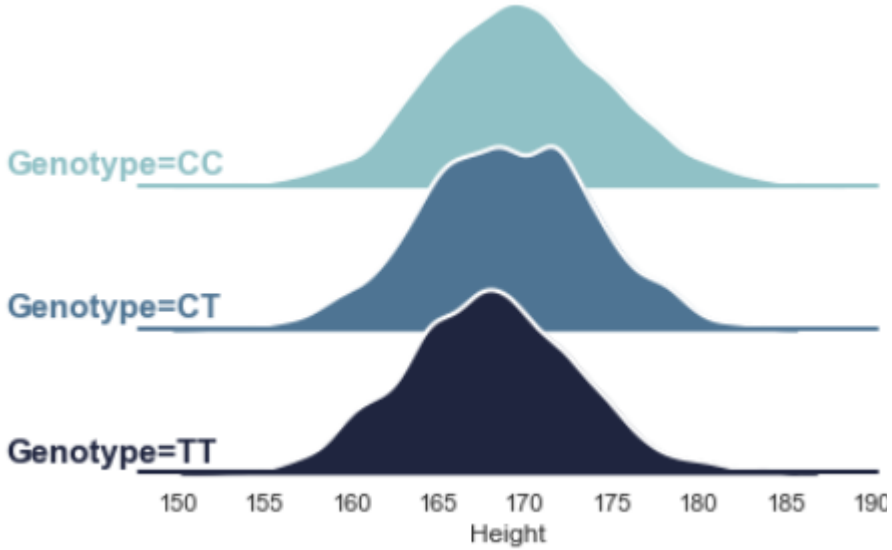
Effect size of 2cm per G allele

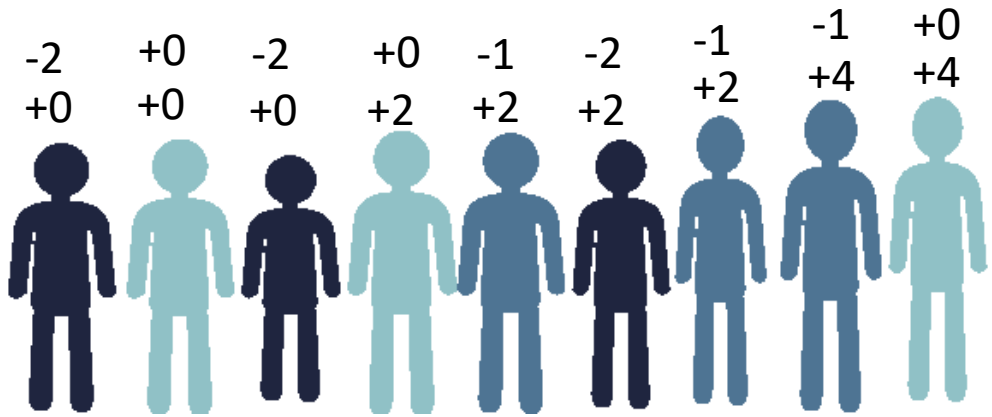






Effect size of -1 per T allele





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

- 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:

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

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

$$\vdots$$

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

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

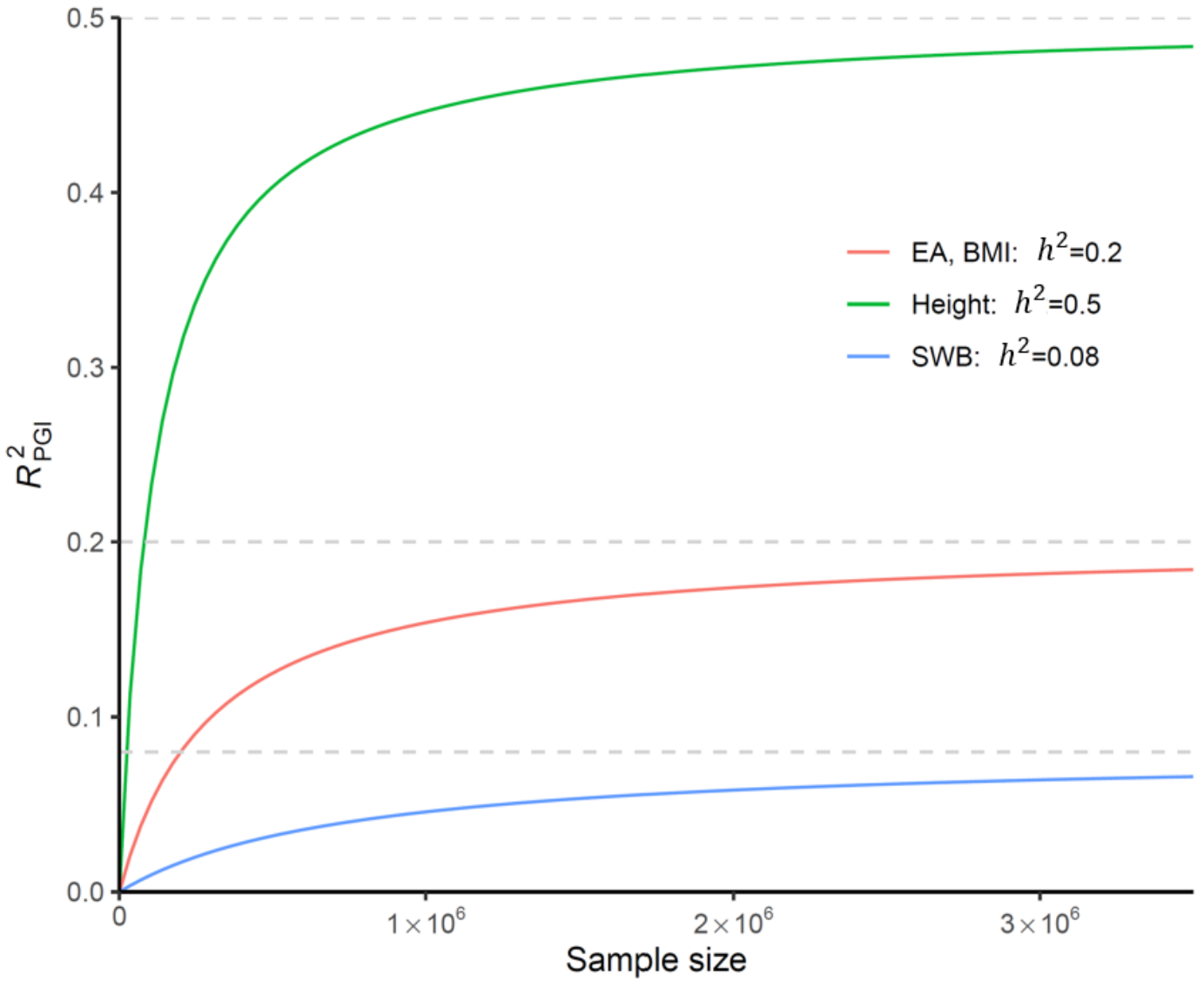
Effective number of SNPs in the PGI, estimated to be between 50k-70k in genome-wide data for EUR ancestry (Wray et al. 2013)

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 we are predicting into a cohort where the genetic architecture is not the same as the GWAS sample?

y, A_{SNP} : phenotype and additive SNP factor in the training (GWAS) sample

y^*, A^*_{SNP} : phenotype and additive SNP factor in the validation sample

$$A^*_{SNP,i} \neq A_{SNP,i} \rightarrow h^{2^*}_{SNP} \equiv \frac{Var(A^*_{SNP,i})}{Var(y^*_i)} \neq h^2_{SNP}$$

Define the genetic correlation to be

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

The expected predictive power

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

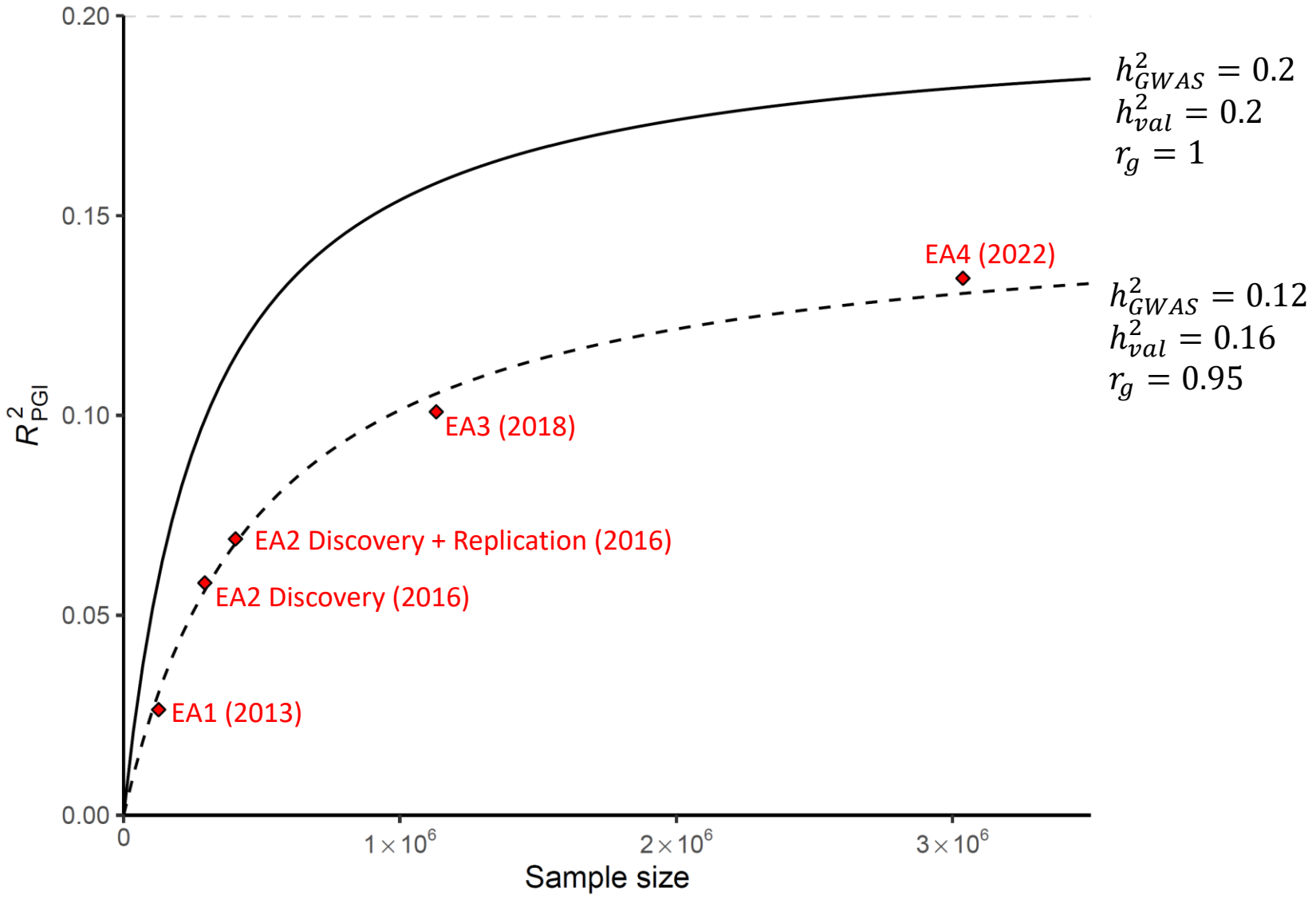
now becomes

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

(De Vlaming et al. 2016)

This formula will hold even if y^*_i is a different phenotype!

Theoretical projections for R^2_{PGI} vs Observed R^2_{PGI}



Outline

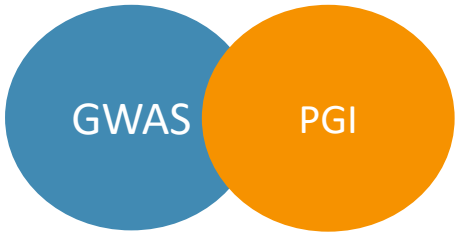
- What is a polygenic index?
- Predictive power of polygenic indices
- Constructing polygenic indices
- Applications
- Limitations & pitfalls

Constructing polygenic indices

What is needed?

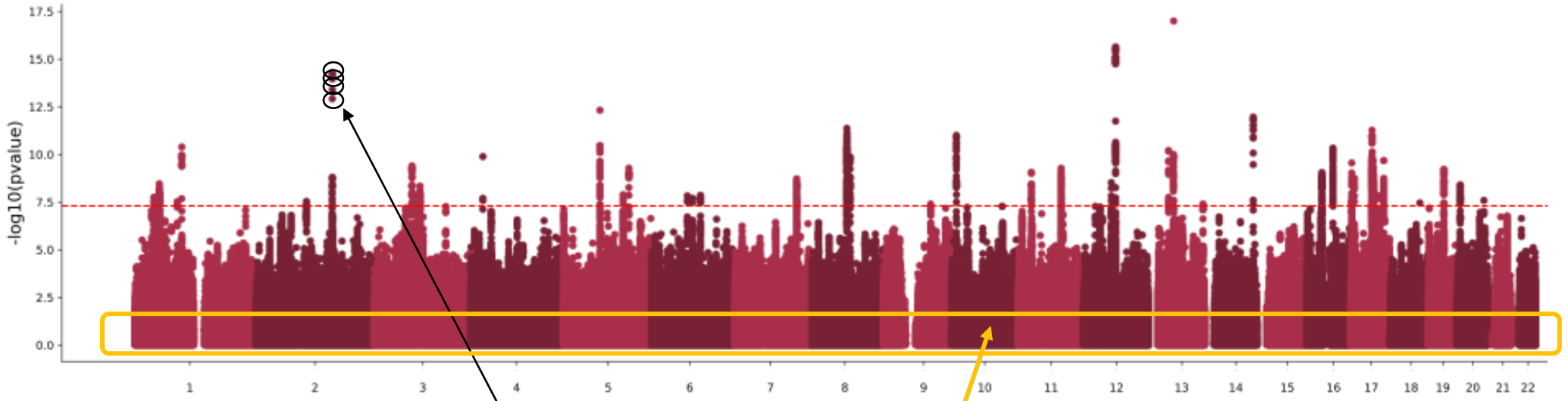
- Individual-level genotype data from a prediction sample.
- Weights: GWAS summary statistics from a discovery sample

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

Clumping and thresholding

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

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

Bayesian approaches

Include all SNPs but adjust the effect sizes for LD

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**

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)

Practical considerations - (C+T)

P-value cutoff: Depends on

- the polygenicity of the trait
  - For highly polygenic traits, reasonable to expect prediction R^2 to increase when more SNPs are included
- the sample size of the discovery GWAS
  - smaller the GWAS sample, the larger the *P*-values → imposing a very strict *P*-value threshold may drop too many SNPs in a small GWAS.

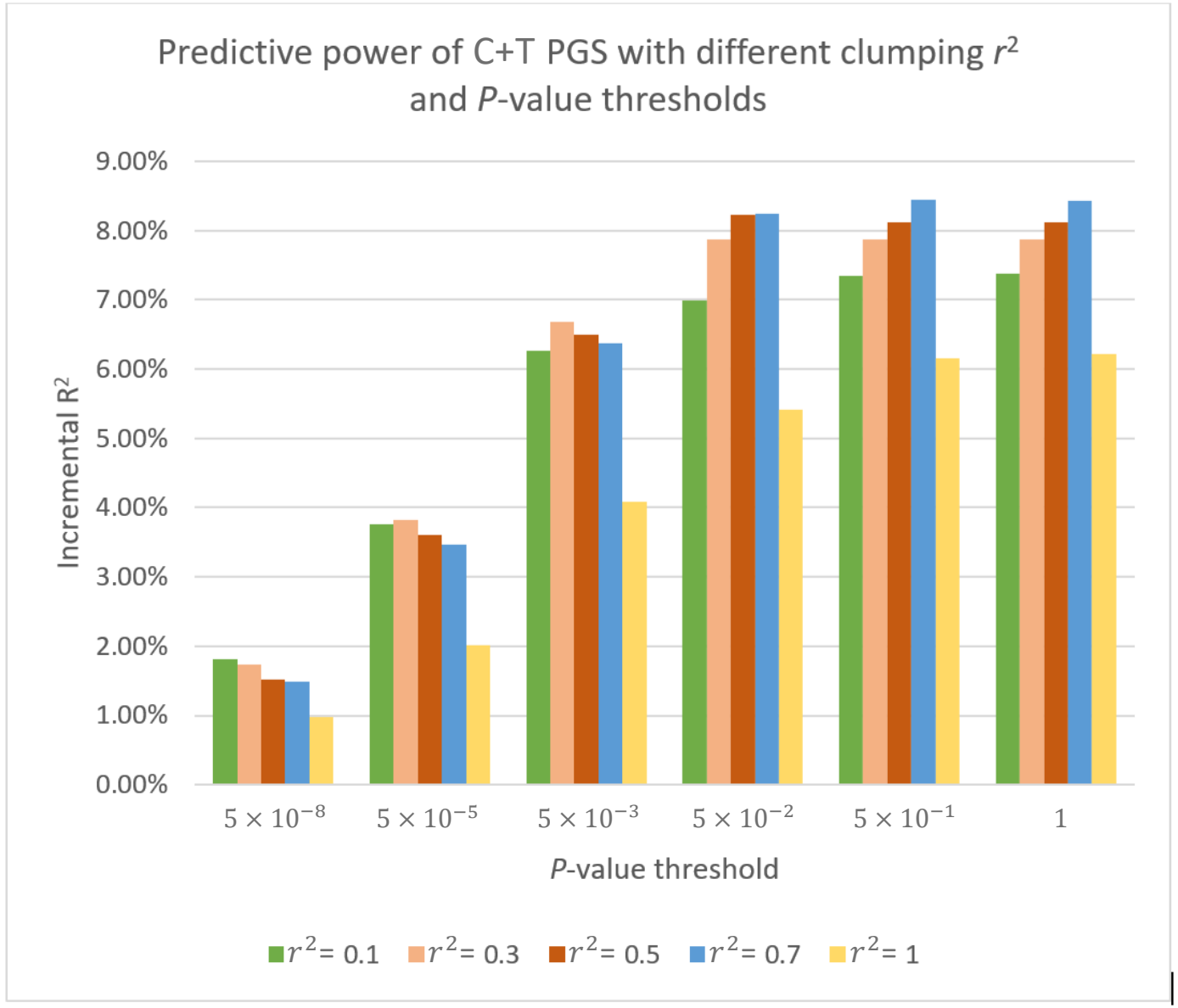
Clumping parameters

- r^2 threshold: Do not want to double-count, but also do not want to lose signal
- LD-window:
  - If too large, then errors in LD estimates can lead to apparent LD between unlinked loci.
  - If too small, there is risk of not accounting for LD between linked loci.

Imputed or genotyped SNPs?

Depends on

- genotyping chip coverage
- quality of imputed SNPs



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

Practical considerations - (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 - (Bayesian approaches)

Imputed or genotyped SNPs?

Same tradeoff between coverage and noise, but

- All SNPs included in the PGI also need to be in the ref genotype data used to calculate the LD matrix
- If using imputed SNPs in the PGI, will either need
  - Imputed reference data to calculate LD → imputation uncertainty introduces noise to LD calculation
  - A large enough and representative sequenced sample → may not be available

The same consideration applies to C+T but Bayesian approaches are more sensitive to noise in LD estimates!

Solutions:

1. Use genotyped SNPs and genotype data from the validation cohort to estimate LD → may not be optimal if
   - the number of genotyped SNPs is low
   - sample size is low
   - cohort has been genotyped using multiple chips
   - want to compare prediction results between different cohorts, and hence need the PGI to include the same set of SNPs
2. Include only SNPs with imputation accuracy above a certain threshold
3. Use HapMap3 SNPs from the imputed data

Practical considerations - (Bayesian approaches)

Reference genotype data to calculate LD matrix should be

- large enough
- representative of the GWAS sample
- 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)

Which method is better?

Clumping and thresholding

Faster and easier, but too black & white

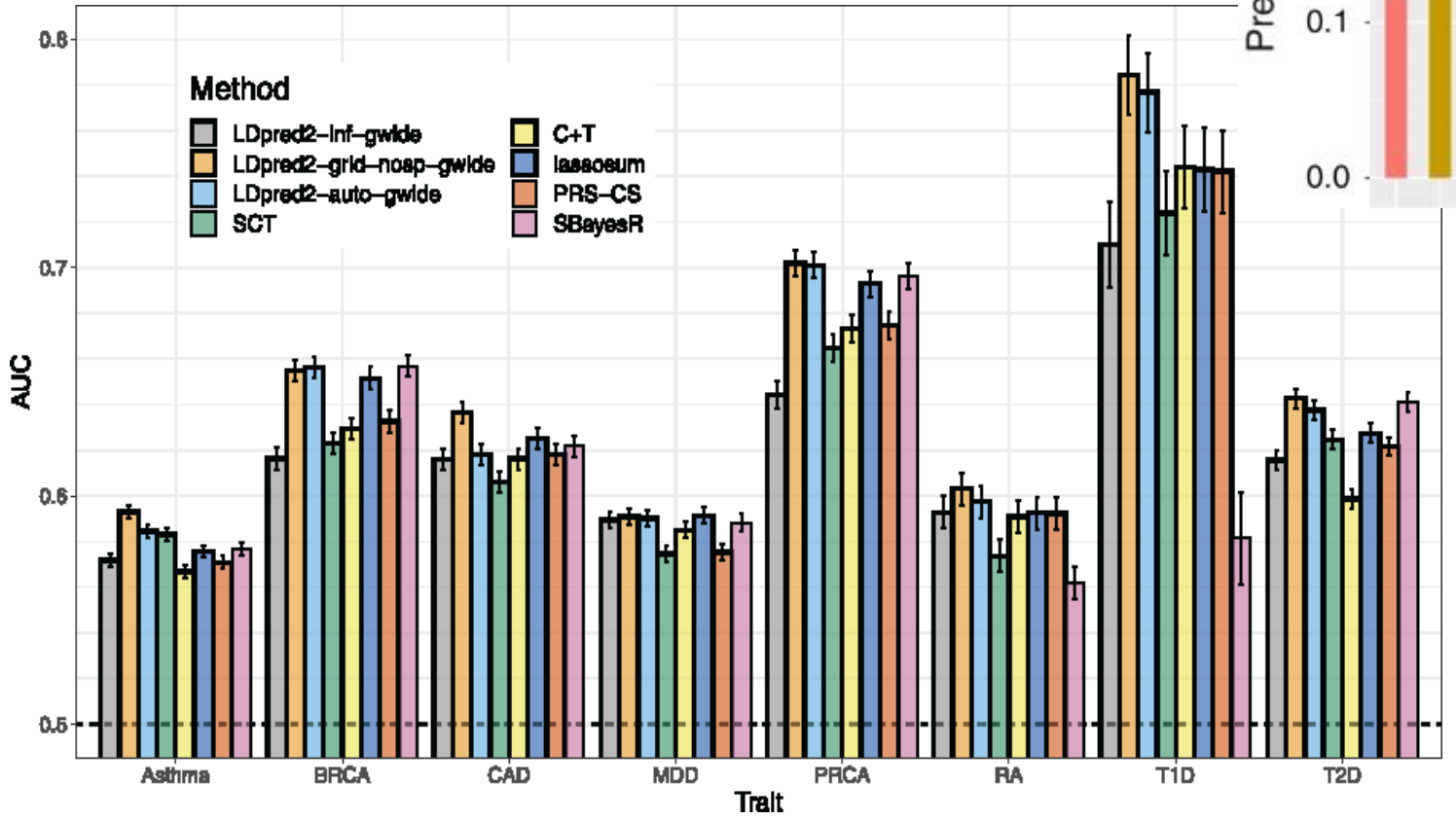
- If clumping r^2 or *P*-value cutoffs too strict, it drops potentially causal SNPs.
- If clumping r^2 and *P*-value cutoffs too relaxed, there is a lot of double-counting and noise

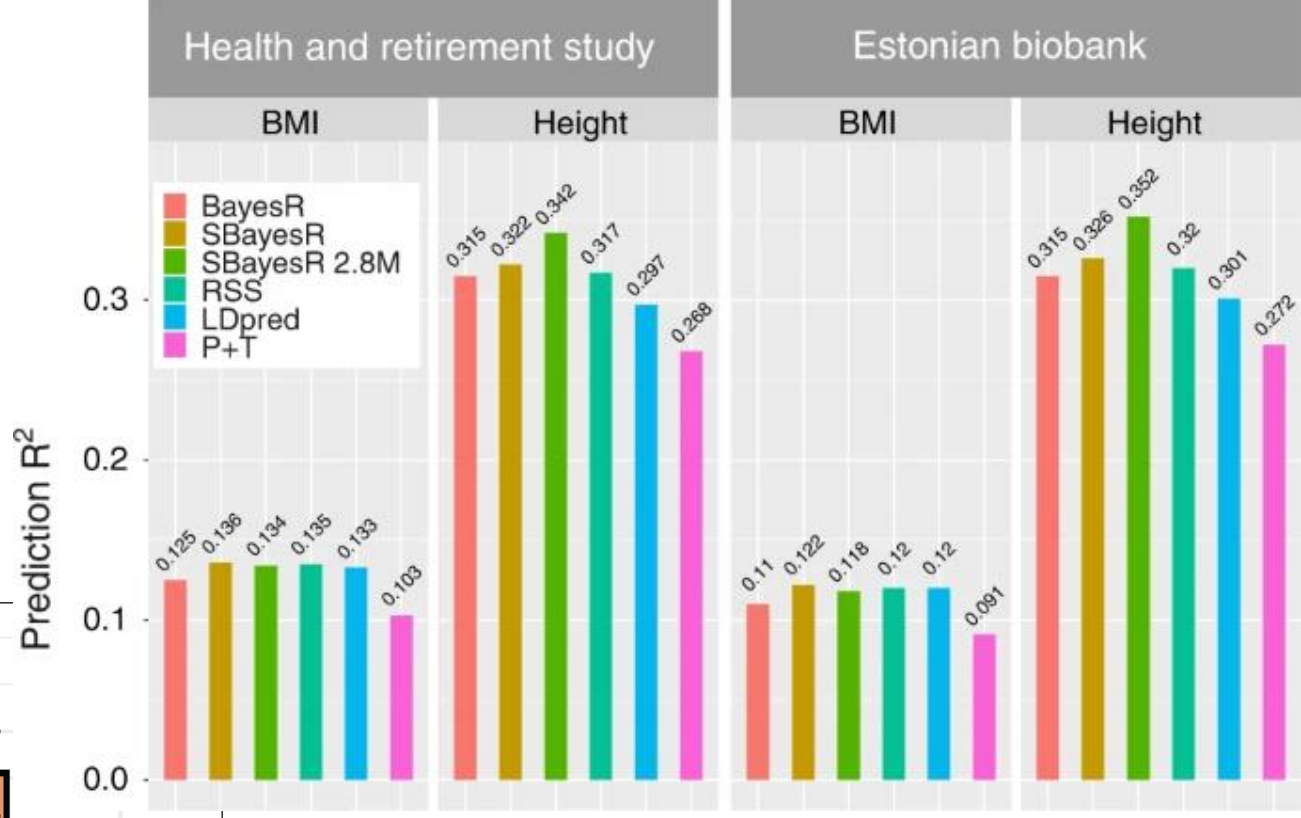
Bayesian approaches

- 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, than Bayesian approaches clearly do better

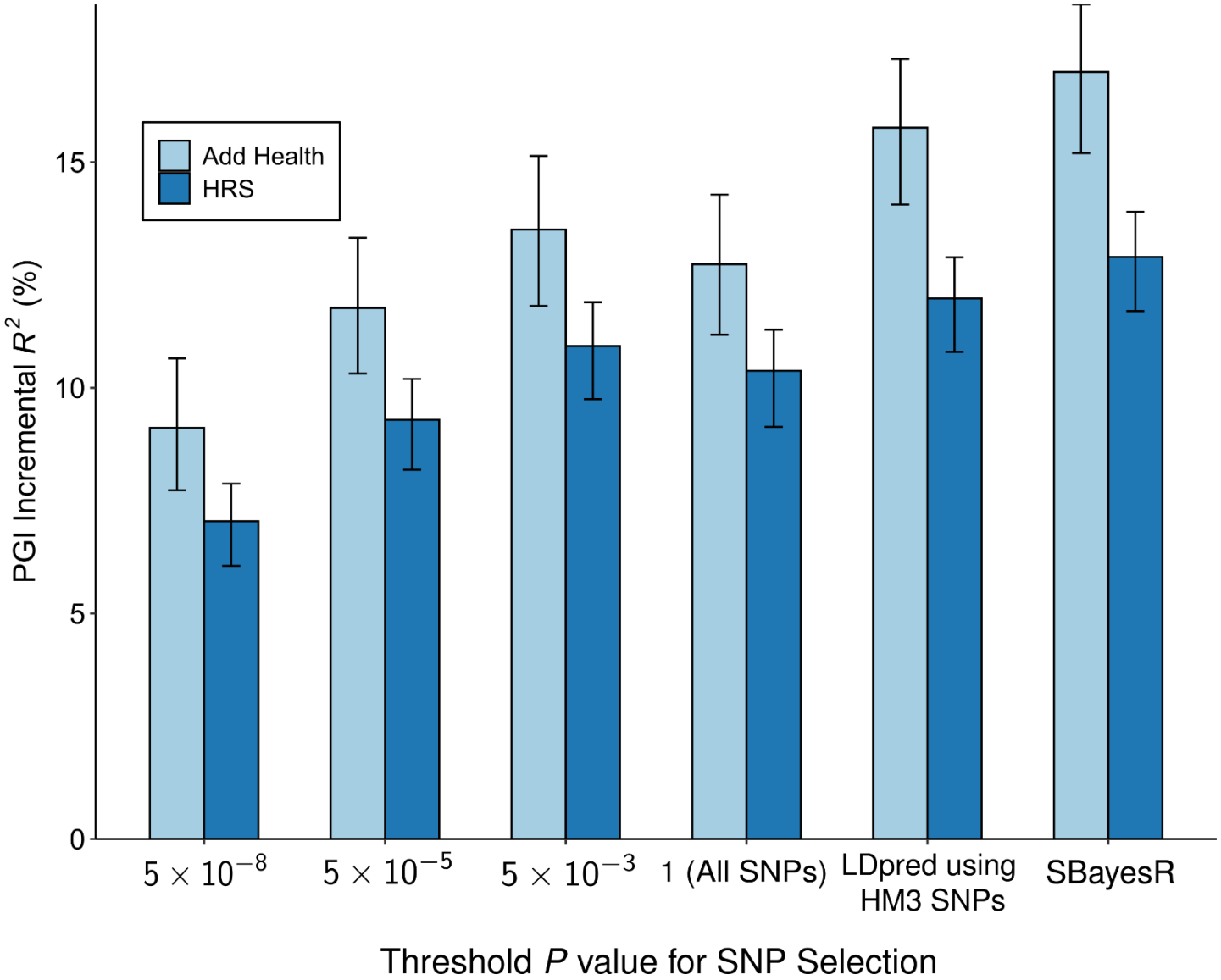
Source: Privé, Arbel, Vilhjálmsson (2020)





Source: Lloyd-Jones et al (2019)

There may still be uses for C+T, e.g. explore how much variance is explained out-of-sample by the genome-wide significant loci



Source: Okbay et al. (2022)

Outline

- 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 treatment effect heterogeneity by genotype**

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)

→ **Use for balance tests of randomization**

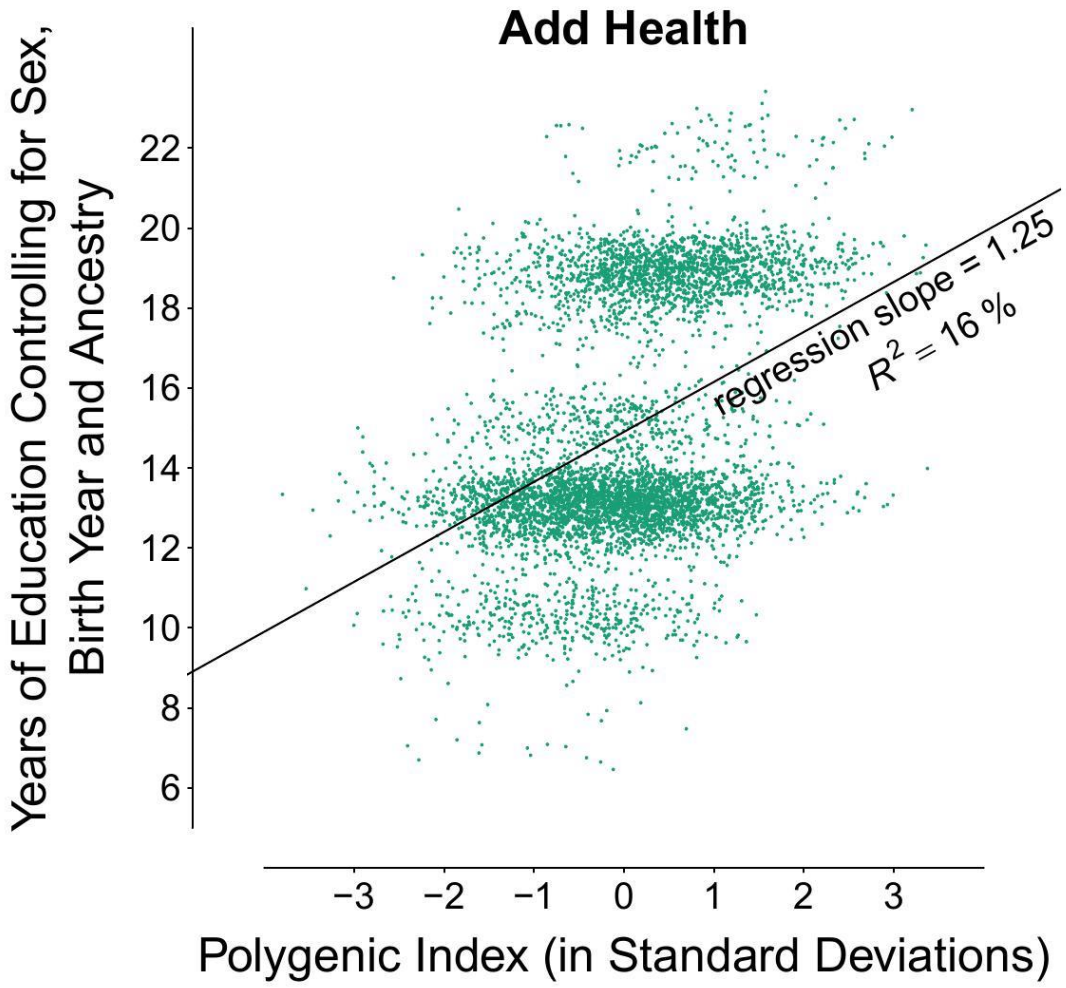
PGIs should be identically distributed in treatment and control groups (Davies et al. 2016, Barcellos, Carvalho, and Turley 2016)

→ **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?

- Can control for the relatedness using the GRM and a linear mixed model
- Possible to do in GCTA
- We will post a video on how to this!

Outline

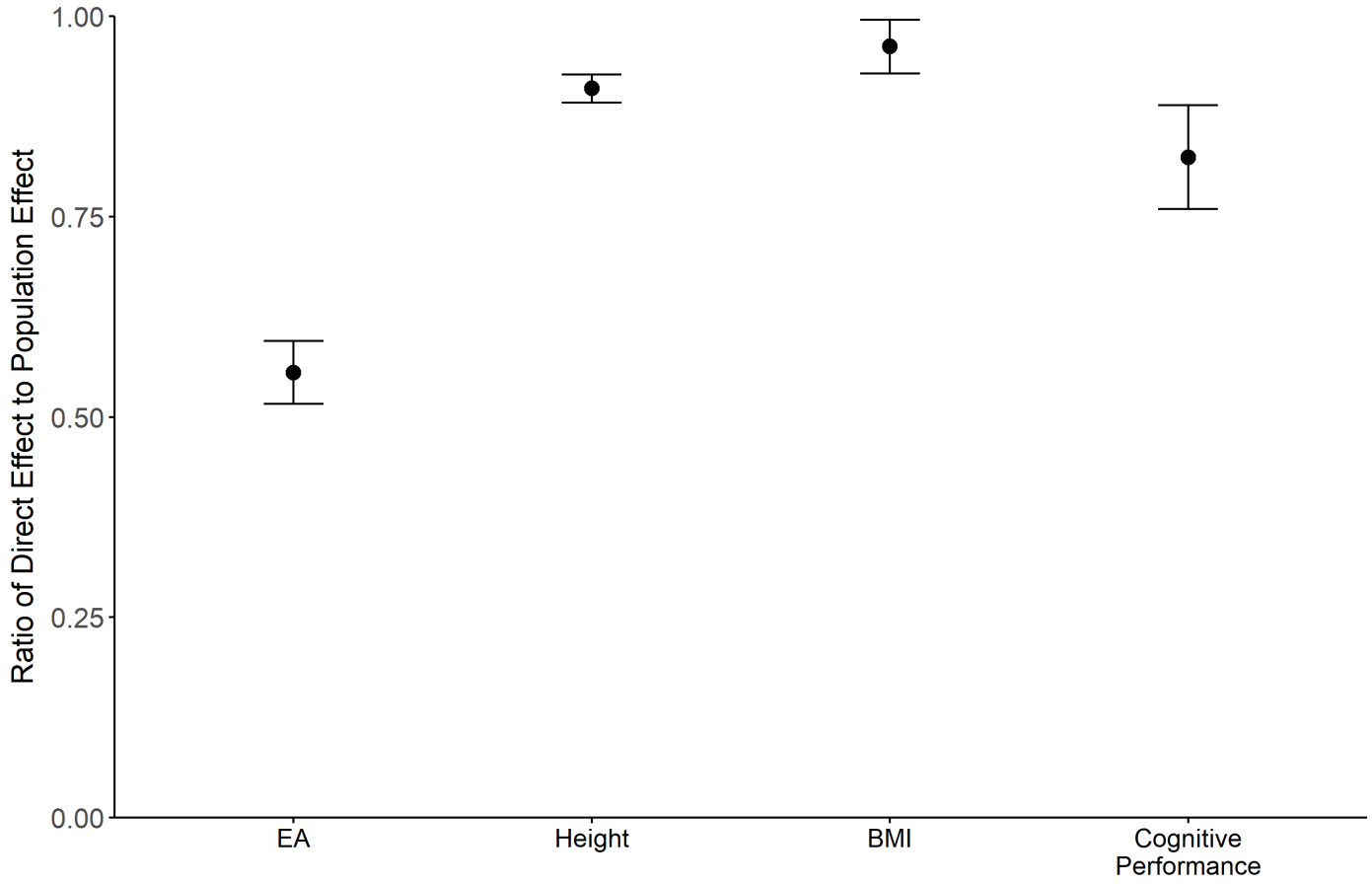
- What is a polygenic index?
- Predictive power of polygenic indices
- Constructing polygenic indices
- Applications
- Limitations and pitfalls

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.

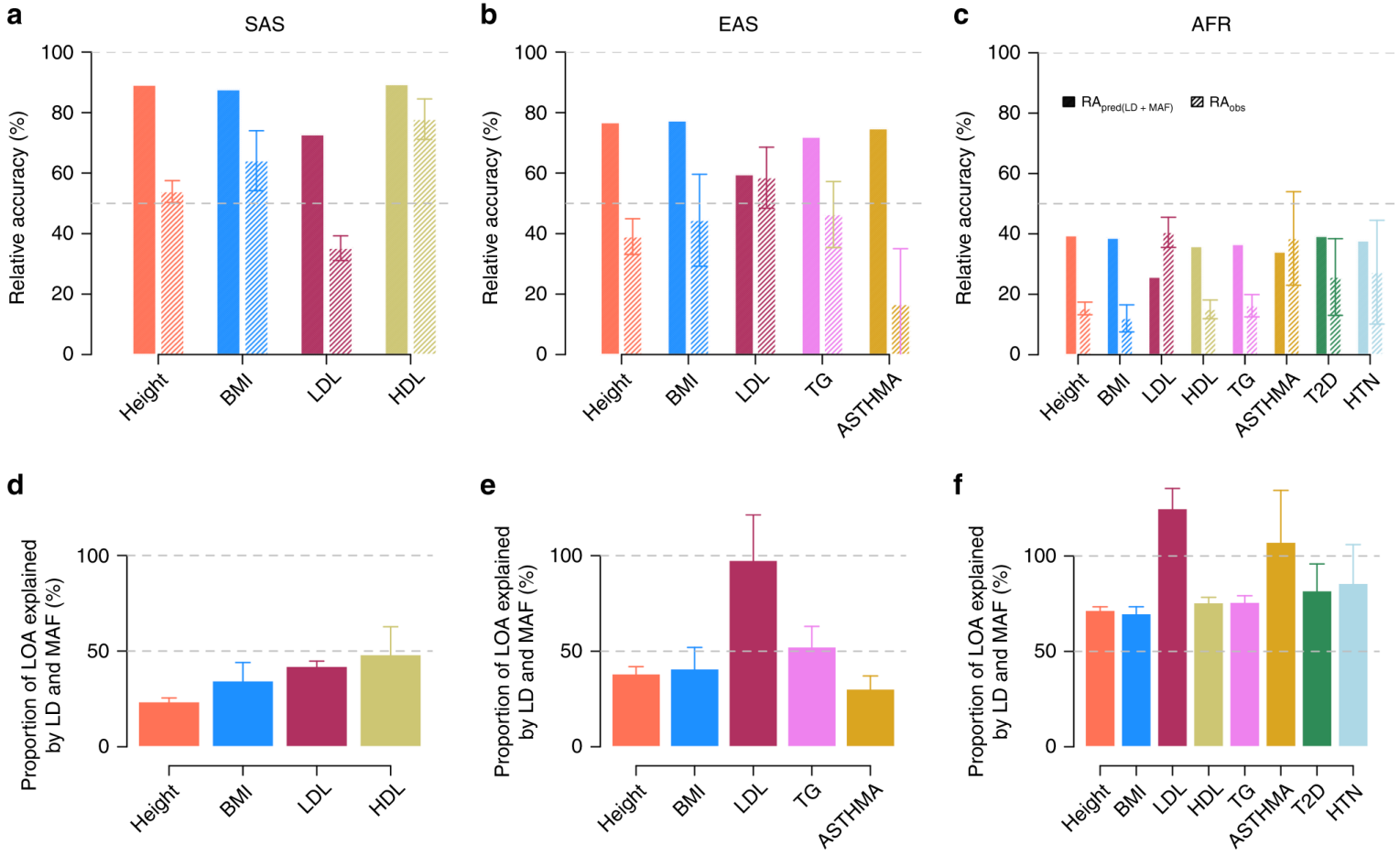


Source: Okbay et al. (2022)

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.

  → Relative accuracies of 15% and 11%



Source: Wang *et al.* (2020)

LIMITATIONS & PITFALLS

Two sources of population stratification

- In the discovery phase
  - leads to bias in the GWAS estimates, so the PGI may give more weight to SNPs that just correspond to ancestry
- In the prediction phase
  - If the prediction sample is stratified, this can lead to bias in our PGI-based analyses even if SNP-weights are unbiased
- Interaction of bias in both phases
  - The combination of these two interact so group differences are strongly exaggerated

→ Important to control for PCs in prediction analyses!

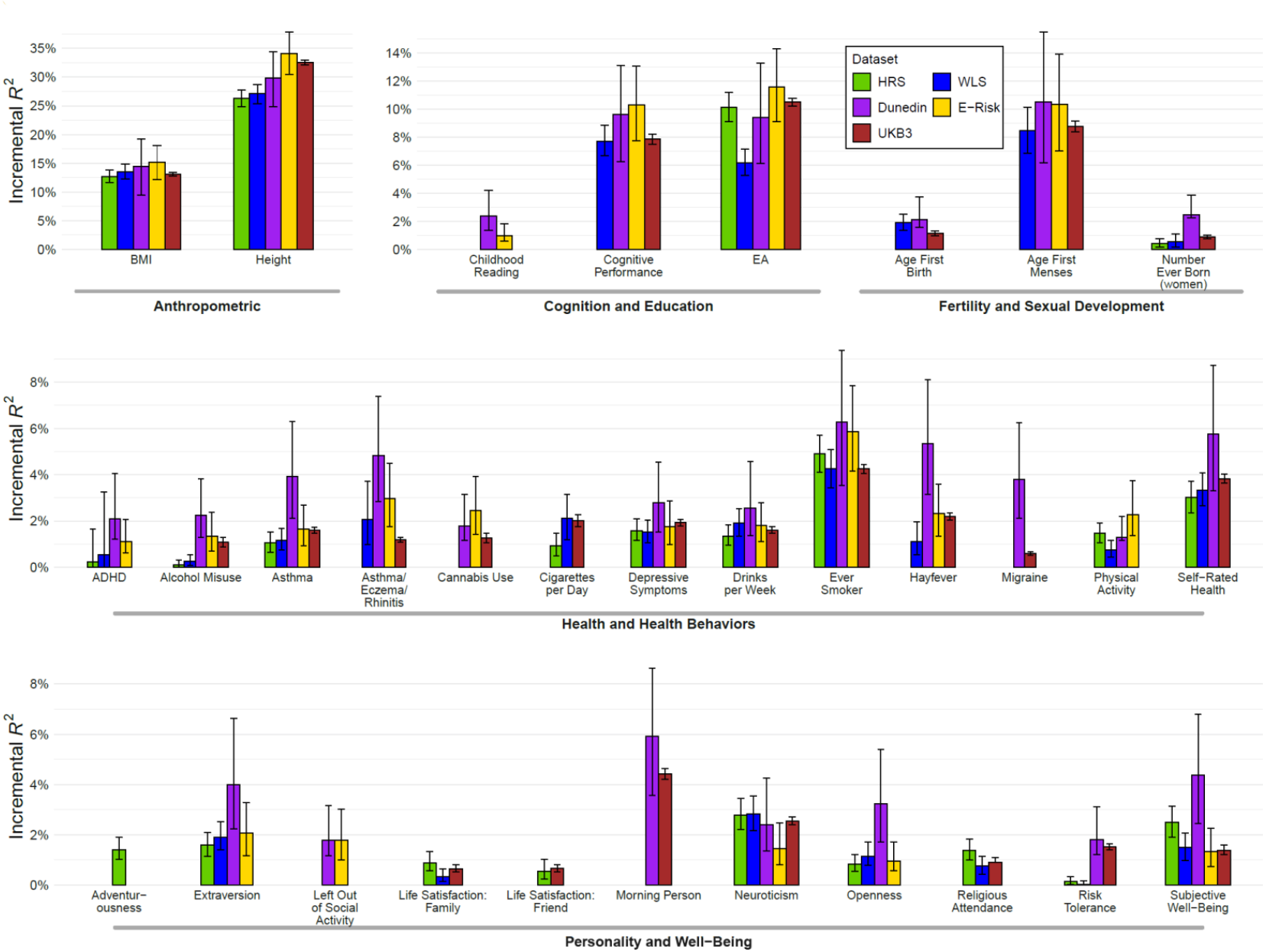
PGI Repository

v1.0

- 47 phenotypes
- 11 cohorts
  - Dunedin Multidisciplinary Health and Development Study
  - English Longitudinal Study of Ageing (ELSA)
  - Environmental Risk (E-Risk) Longitudinal Twin Study
  - Estonian Biobank
  - Health and Retirement Study
  - Minnesota Center for Twin and Family Research (MCTFR)
  - National Longitudinal Study of Adolescent to Adult Health
  - Swedish Twin Registry
  - Texas Twin Project
  - UK Biobank
  - Wisconsin Longitudinal Study

v2.0 (coming soon)

- 7 new cohorts, 21 new phenotypes
- Parental PGIs



Source: Becker et al. (2021)

QUESTIONS?

References

- Becker, J., Burik, C. A. P. P., Goldman, G., Wang, N., Jayashankar, H., Bennett, M., Belsky, D. W., Karlsson Linnér, R., Ahlskog, R., Kleinman, A., Hinds, D. A., Agee, M., Alipanahi, B., Auton, A., Bell, R. K., Bryc, K., Elson, S. L., Fontanillas, P., Furlotte, N. A., … Okbay, A. (2021). Resource profile and user guide of the Polygenic Index Repository. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-021-01119-3
- Daetwyler, H. D., Villanueva, B., & Woolliams, J. A. (2008). Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. *PLoS ONE*, *3*(10), e3395. https://doi.org/10.1371/journal.pone.0003395
- de Vlaming, R., Okbay, A., Rietveld, C. C. A., Johannesson, M., Magnusson, P. K. E., Uitterlinden, A. G., van Rooij, F. J. A., Hofman, A., Groenen, P. J. F., Thurik, A. R. R., Koellinger, P. D. P., Wood, A., Esko, T., Yang, J., Vedantam, S., Pers, T., Gustafsson, S., Locke, A., Kahali, B., … Ritchie, S. (2017). Meta-GWAS Accuracy and Power (MetaGAP) Calculator Shows that Hiding Heritability Is Partially Due to Imperfect Genetic Correlations across Studies. *PLOS Genetics*, *13*(1), e1006495. https://doi.org/10.1101/048322
- Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., & Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, *10*(1), 1776. https://doi.org/10.1038/s41467-019-09718-5
- Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., Wang, H., Zheng, Z., Magi, R., Esko, T., Metspalu, A., Wray, N. R., Goddard, M. E., Yang, J., & Visscher, P. M. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nature Communications*, *10*(1), 5086. https://doi.org/10.1038/s41467-019-12653-0
- Okbay, A., Wu, Y., Wang, N., Jayashankar, H., Bennett, M., Nehzati, S. M., Sidorenko, J., Kweon, H., Goldman, G., Gjorgjieva, T., Jiang, Y., Hicks, B., Tian, C., Hinds, D. A., Ahlskog, R., Magnusson, P. K. E., Oskarsson, S., Hayward, C., Campbell, A., … Young, A. I. (2022). Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics 2022 54:4*, *54*(4), 437–449. https://doi.org/10.1038/s41588-022-01016-z
- Privé, F., Luu, K., Blum, M. G. B., McGrath, J. J., & Vilhjálmsson, B. J. (2020). Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics*, *36*(16), 4449–4457. https://doi.org/10.1093/bioinformatics/btaa520
- Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., Sklar, P., McQuillin, A., Morris, D. W., O'Dushlaine, C. T., Corvin, A., Holmans, P. A., MacGregor, S., Gurling, H., Blackwood, D. H. R. R., Craddock, N. J., Gill, M., Hultman, C. M., Kirov, G. K., … Consortium, T. I. S. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748–752. https://doi.org/10.1038/nature08185
- Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P. R., Bhatia, G., Do, R., Hayeck, T., Won, H. H., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K. H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., … Price, A. L. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *American Journal of Human Genetics*, *97*(4), 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001
- Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P. M., & Yengo, L. (2020). Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nature Communications*, *11*(1), 1–9. https://doi.org/10.1038/s41467-020-17719-y
- Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., & Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, *14*(7), 507–515. https://doi.org/10.1038/nrg3457

Further reading

- Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, *9*(3), e1003348. https://doi.org/10.1371/journal.pgen.1003348
- Evans, D. M., Visscher, P. M., & Wray, N. R. (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human Molecular Genetics*, *18*(18), 3525–3531. https://doi.org/10.1093/hmg/ddp295
- Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, *51*(4), 584–591. https://doi.org/10.1038/s41588-019-0379-x
- Witte, J. S., Visscher, P. M., & Wray, N. R. (2014). The contribution of genetic variants to disease depends on the ruler. *Nature Reviews Genetics*, *15*(11), 765–776. https://doi.org/10.1038/nrg3786
- Wray, N. R., Goddard, M. E., & Visscher, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research*, *17*(10), 1520–1528. https://doi.org/10.1101/gr.6665407
- Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A. E., Dudbridge, F., & Middeldorp, C. M. (2014). Research Review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, *55*(10), 1068–1087. https://doi.org/10.1111/jcpp.12295

PRACTICAL

- Clumping and thresholding PGI
- Obtaining the incremental-R^2
- Confidence intervals for incremental-R^2
- Plotting the results

https://ucsas.qualtrics.com/jfe/form/SV_0xO9zBVxPeJVWZ0