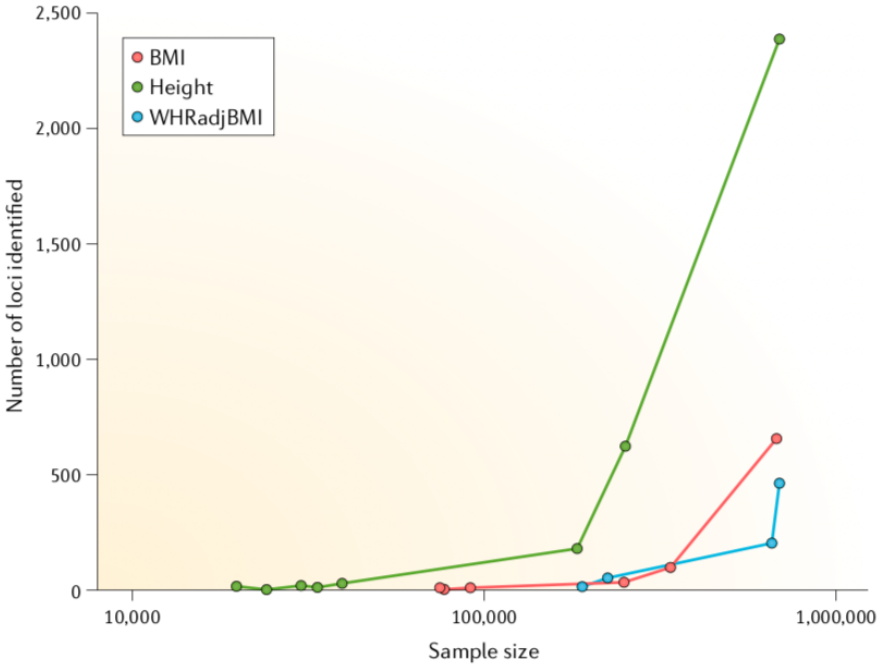
# What drives the reduced prediction accuracy of polygenic scores in non-European individuals?

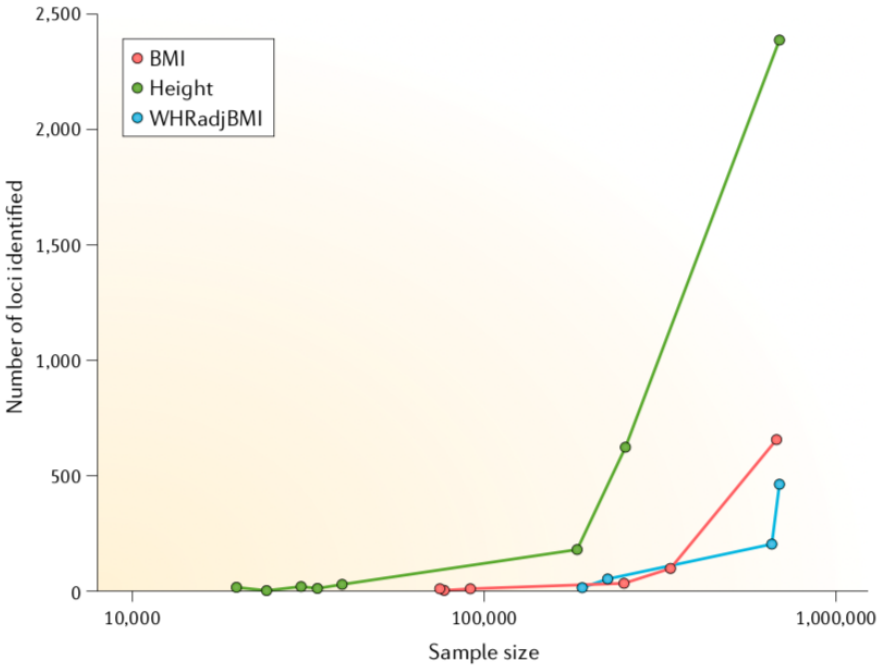
Bárbara Domingues Bitarello

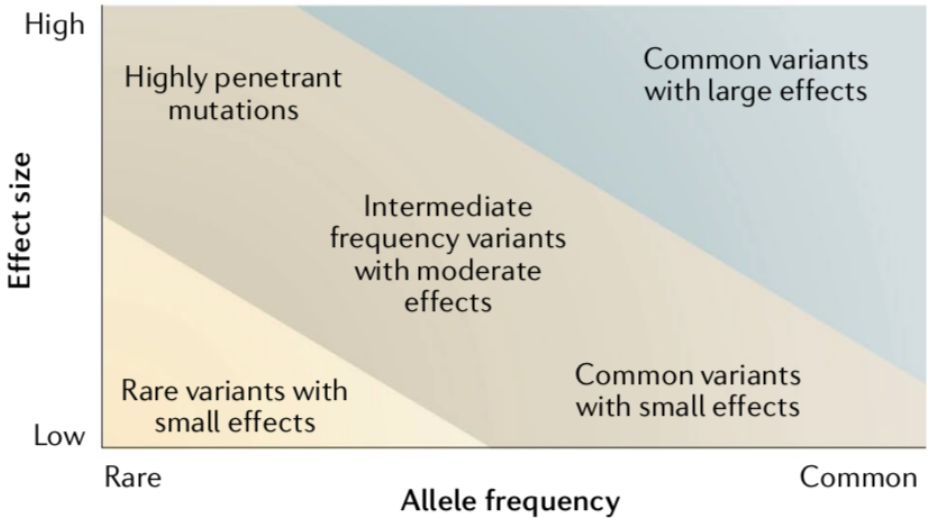
Perelman School of Medicine, University of Pennsylvania

# Many variants



# Many variants with small effect size




[Tam et al. (2019) *Nat Rev Genet*]

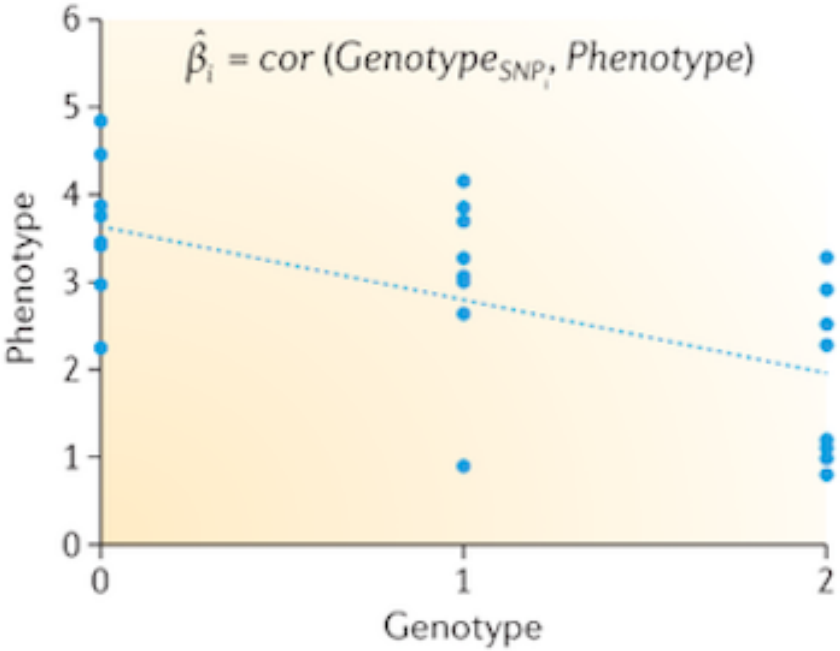
Combined, they explain a lot!

# Some examples

| Phenotype | Statistic | Value | Variants |
|---|---|---|---|
| height | R-squared | 25.0 | 3000 |
| schizophrenia | R-squared | 7.0 | 100 |
| ADHD | R-squared | 5.5 | 100 |
| breast cancer | AUC | 60.0 | 1000 |
| cardiovascular disease (CAD) | AUC | 81.0 | 6000 |

## PS: In Europeans...

# Polygenic risk scores add up those small effects




Pasaniuc & Price (2017), *Nat Rev Genet*

$$PRS = \sum_{i=1}^{m} \hat{\beta}_i G_{j,i}$$

$\hat{\beta}$ : effect size (from GWAS)

$G$ : Effect allele dosage

$j$: Individuals

$i$: SNPs

independence

additive model

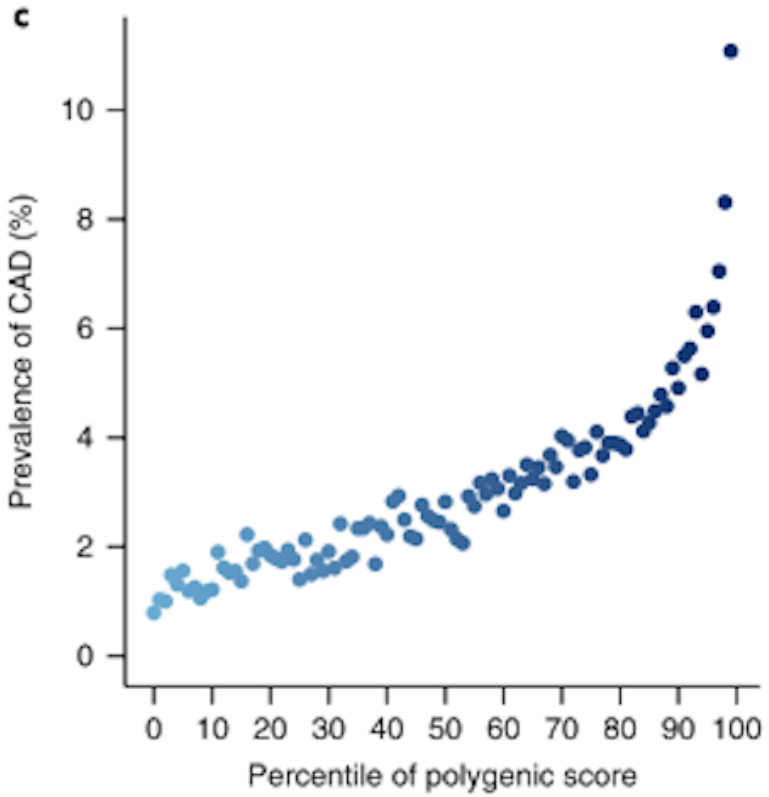
# PRSs are appealing

easy

promising

fast

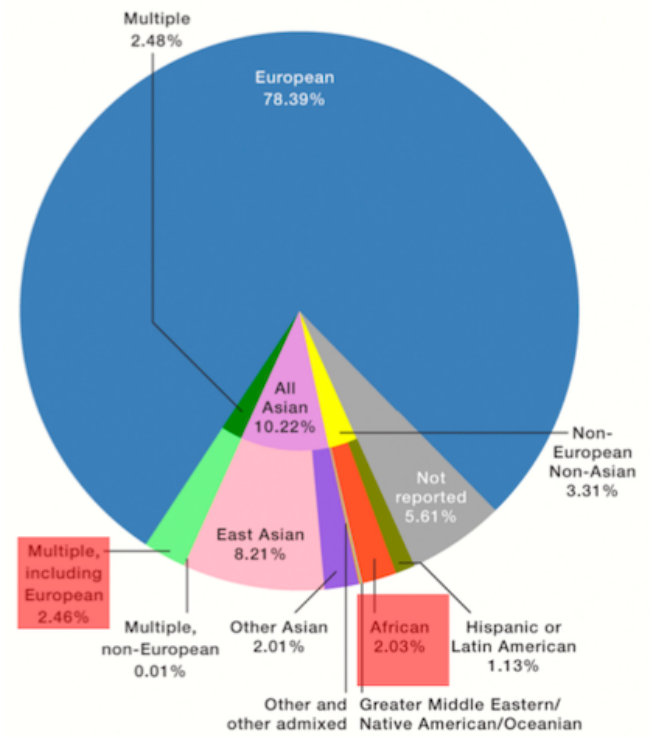
minimal requirements



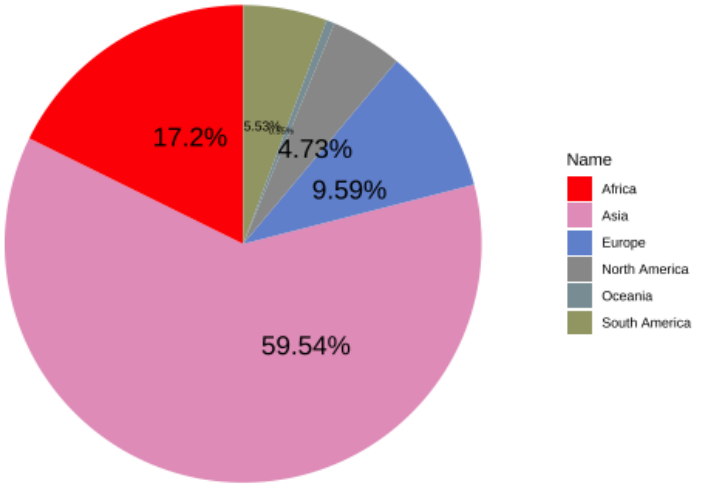

[Khera et al (2018) Nat Genet]

# What about ancestry?

# Europeans represent almost 80% of GWAS participants...




[Sirugo, Williams & Tishkoff (2019) *Cell*]

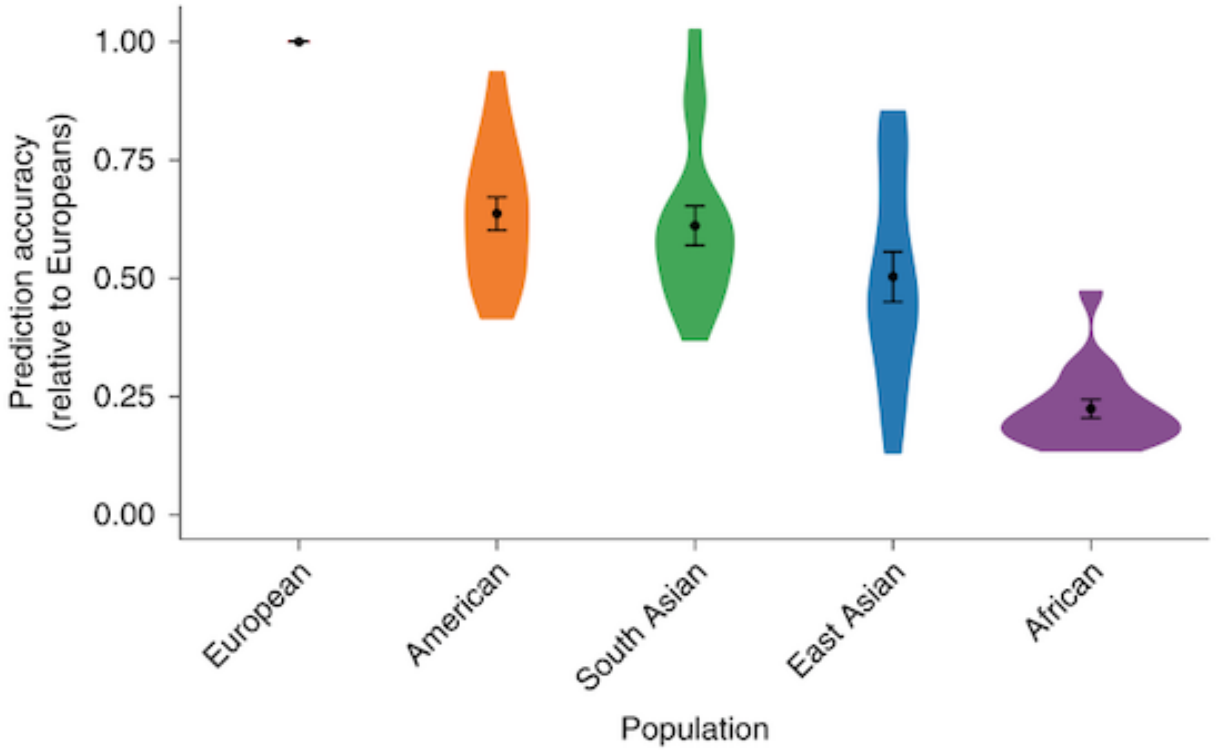



[Data: https://worldpopulationreview.com/]

# .. and 10% of the world's population

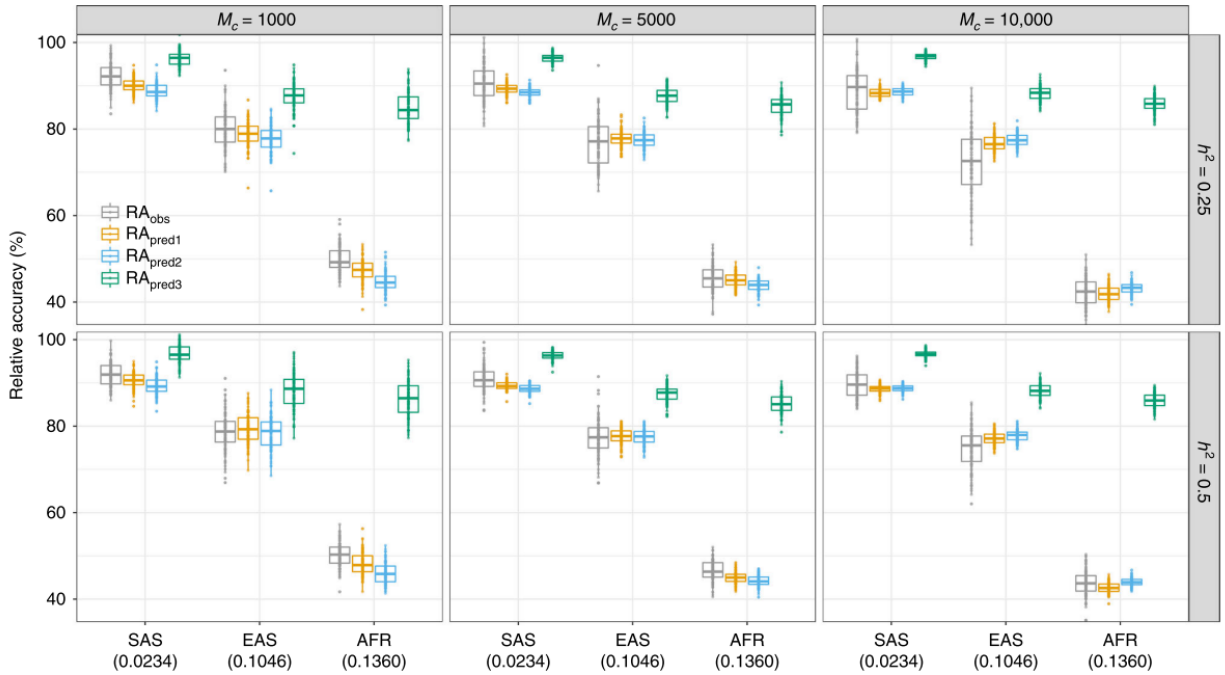
# How does PRS accuracy transfer across ancestries?

# PRS accuracy decreases with genetic distance from Europeans




[Martin et al. (2019) *Nat Genet*]

# PRS accuracy decreases with genetic distance from Europeans




[Wang & Wisscher (2020) *Nat Comms*]

Why is this loss of prediction observed?

What **can** ***we do*** about it?

# Many factors may impact prediction accuracy

causal variants

local selection

gene-gene interactions

gene-environment interactions

marginal effect sizes

LD

site frequency spectrum

phenotypic variance

**These factors are not mutually exclusive!**

# Questions

How do these different factors affect prediction accuracy?

Can we leverage ancestry into the PRS and improve predictions?

# Let's look at height

# Ancestry as a continuous variable

highly polygenic

many cohorts phenotyped

large GWAS (UKBB

$\sim$ 360,000 Europeans)

$$height \sim Sex + Age + Age^2 + p_{eur}$$

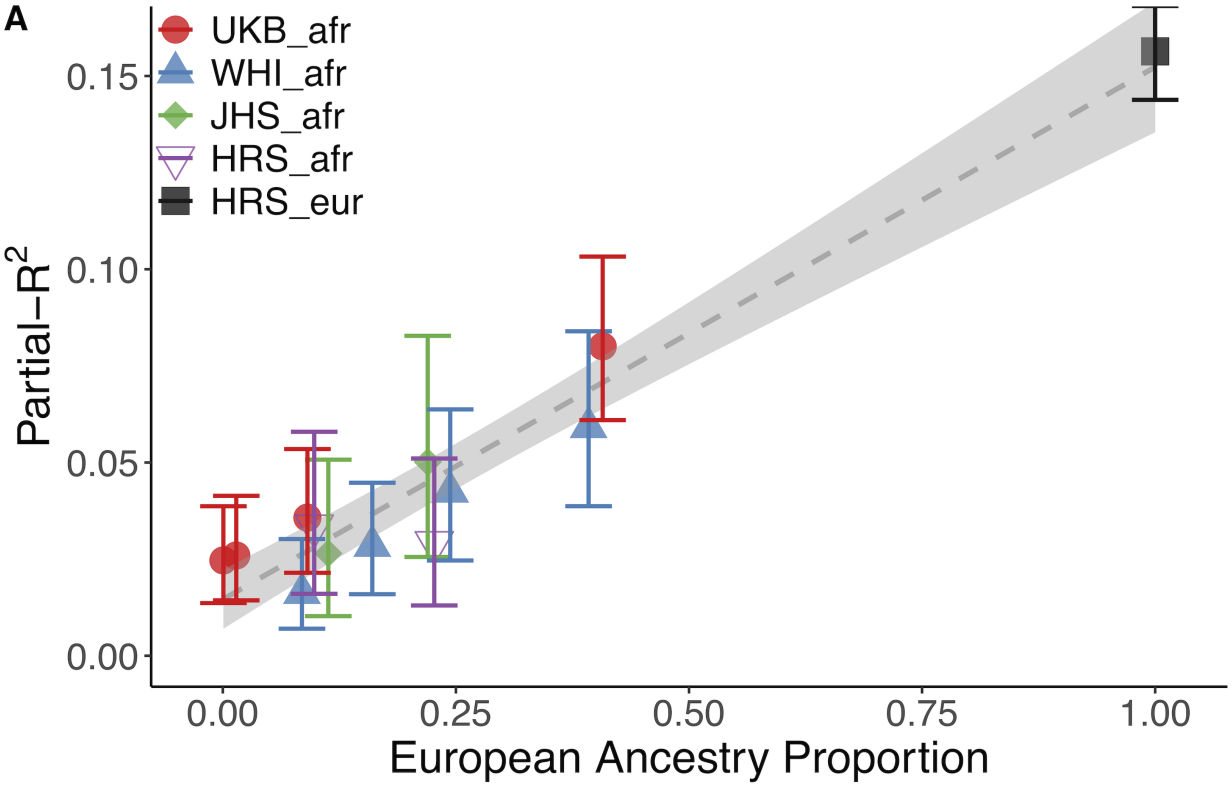
$$height \sim Sex + Age + Age^2 + p_{eur} + PRS$$

European + African ancestry

| Data | Ancestry | N | Number_SNPs |
|---|---|---|---|
| UKBB_eur | European | 9998 | 685475 |
| HRS_EUR | European | 10159 | 1511742 |
| UKBB_afr | African + European | 8700 | 685475 |
| WHI_afr | African American | 6863 | 741983 |
| JHS_afr | African American | 1773 | 702685 |
| HRS_afr | African American | 2251 | 1511742 |

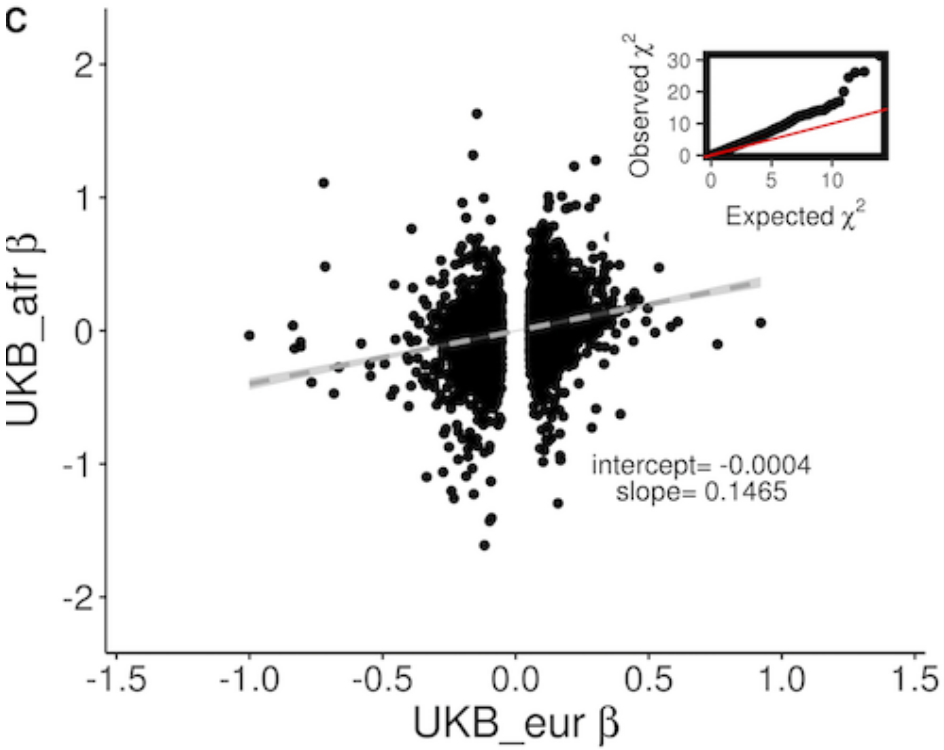
Bitarello & Mathieson (2020), *G3*

# PRS accuracy decreases with proportion of European ancestry




Bitarello & Mathieson (2020), *G3*

# GWAS from UKBB (AFR) individuals




$$y = sex + age + age^2 + 10PCs$$

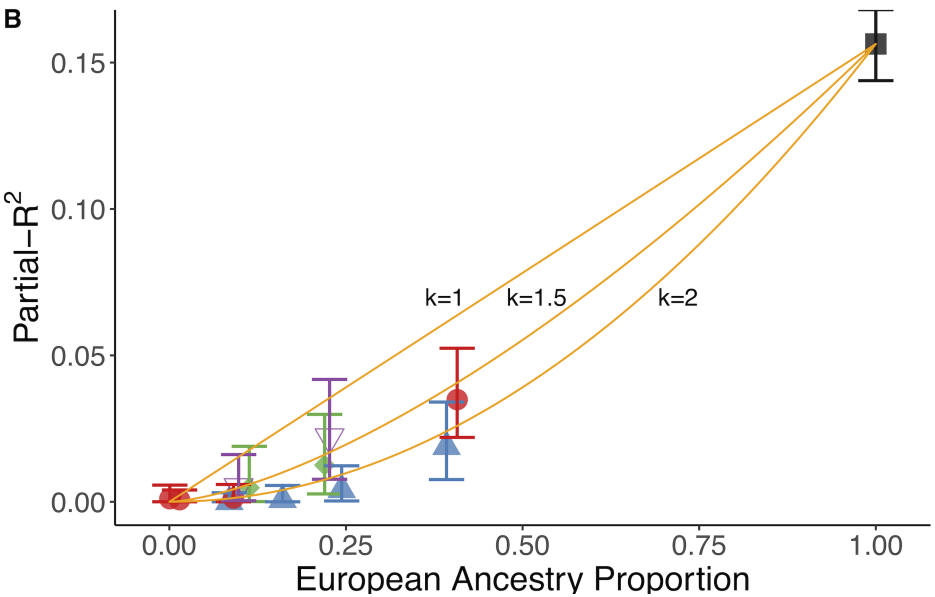
$$N_{AFR} \sim 8,800$$

$$N_{EUR} \sim 350,000$$

$$\chi^2_{diff} = \left[ \frac{\beta_{eur} - \beta_{afr}}{\sqrt{SE^2_{eur} + SE^2_{afr}}} \right]^2$$

Bitarello & Mathieson (2020), *G3*

# Using only EUR chunks of each genome




Bitarello & Mathieson (2020), *G3*

$$y = 0.15^k$$

$$k = 1$$

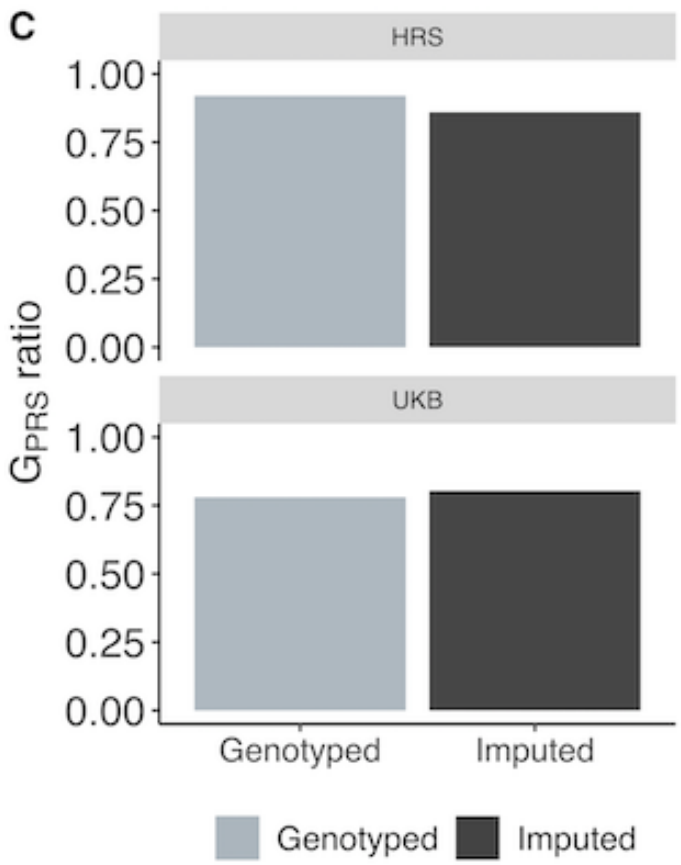
all prediction power comes from European chunks

$$\kappa = 2$$

prediction power is randomly scattered across the genome

**Surprisingly, this relationship seems to be more like $\kappa = 2$**

# Allelic Frequency differences explain about 20%




Additive genetic variance

$$G_{PRS} = \frac{\sum 2f_{i_afr}(1-f_{i,afr})\beta^2_{i,eur}}{\sum 2f_{i,eur}(1-f_{i,eur})\beta^2_{i,eur}}$$

Bitarello & Mathieson (2020), *G3*

# Prediction: phenotypic variance lower in EUR

genome-wide genetic variance in EUR is $\sim 76\%$ of that in AFR

$$height \sim Sex + Age + Age^2 + p_{eur}$$

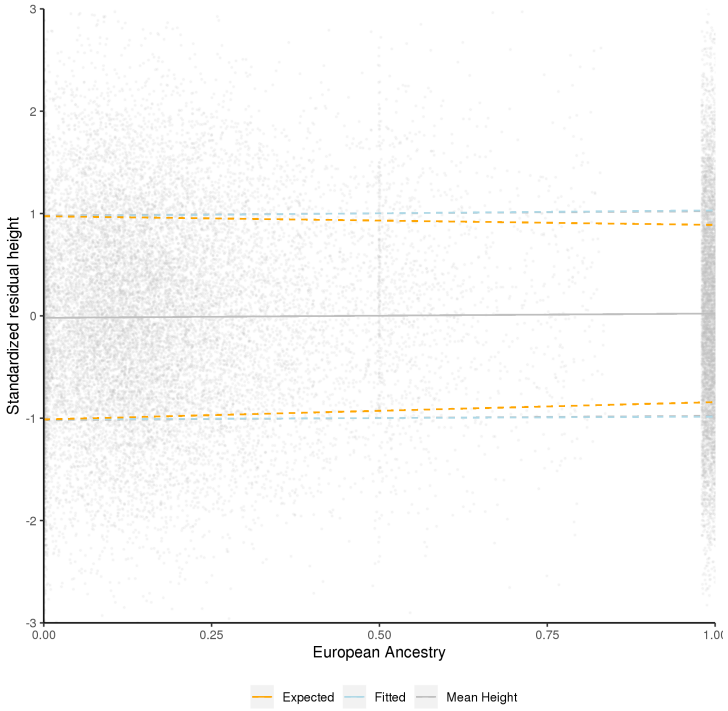
Variable variance model:

$$y = \mu + \beta p_{j,eur} + \epsilon; \epsilon_j \sim N(0, \delta^2 + \gamma p_{j,eur})$$

Mean+1 1sd, constant variance

fitted, variable variance

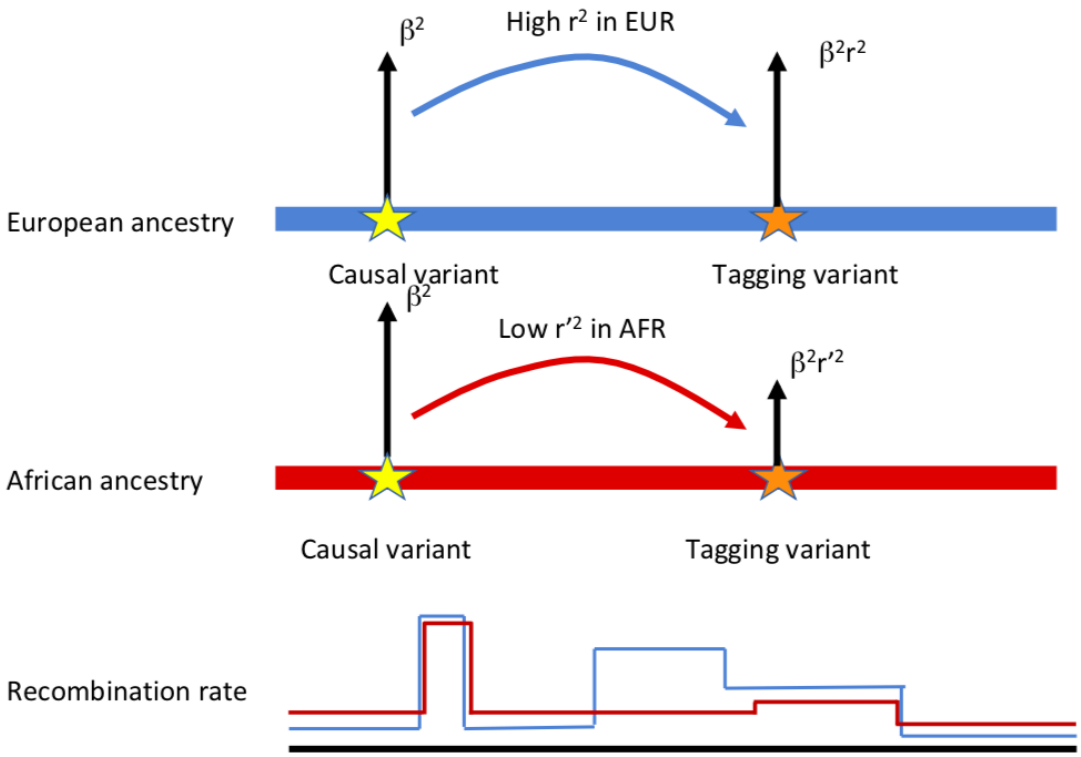
model, variance in phenotypic variance is 100% in AFR and 76% in EUR



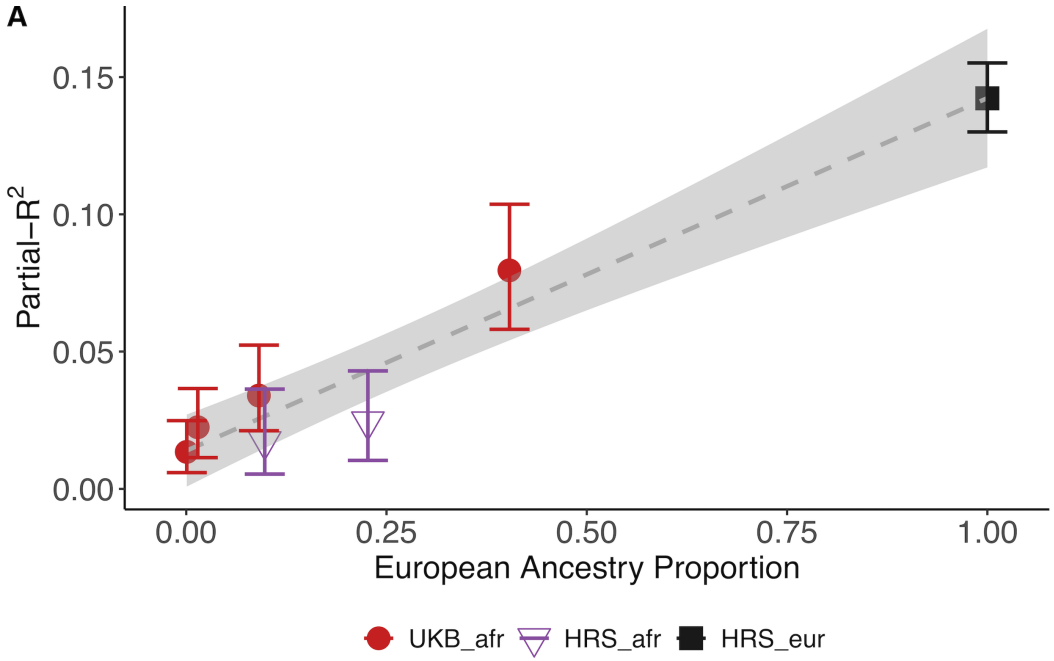

[Bitarello & Mathieson (2020), *G3*]

# Observation: Phenotypic variance does not change with ancestry

# Differences in linkage disequilibrium



## Prediction: better tagging of causal variants decreases loss-of-accuracy

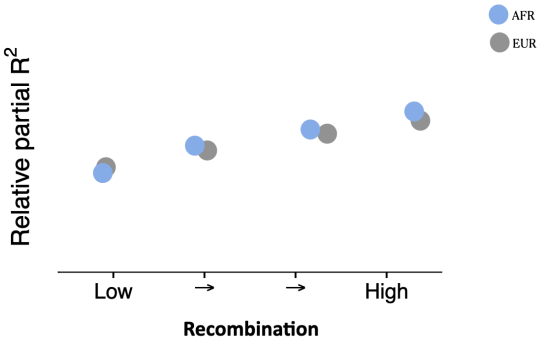


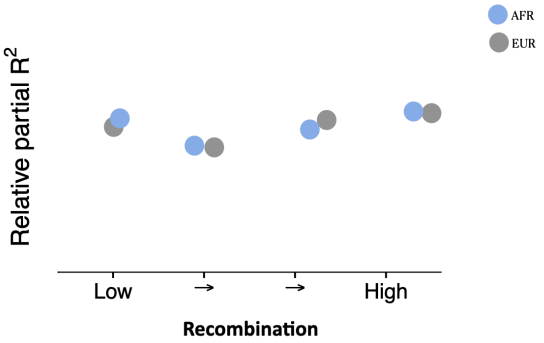

[Bitarello & Mathieson (2020), *G3*]

## Observation: Imputation doesn't help

# Prediction 1: LOA in Africans is independent of LD differences

$$Rel_{R2} = \frac{R^2_{bin}}{R^2_{total}}$$

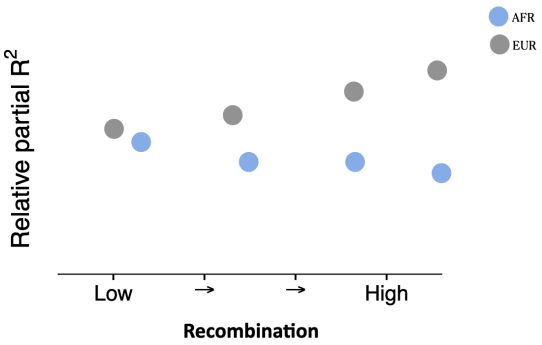


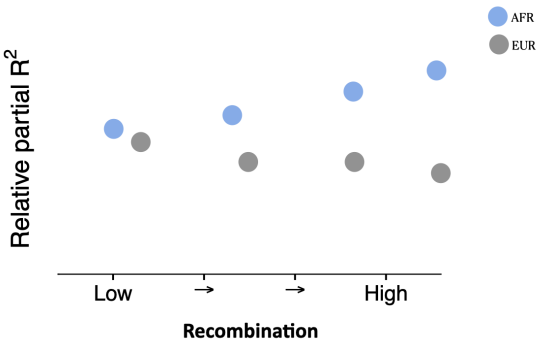




Similar slopes

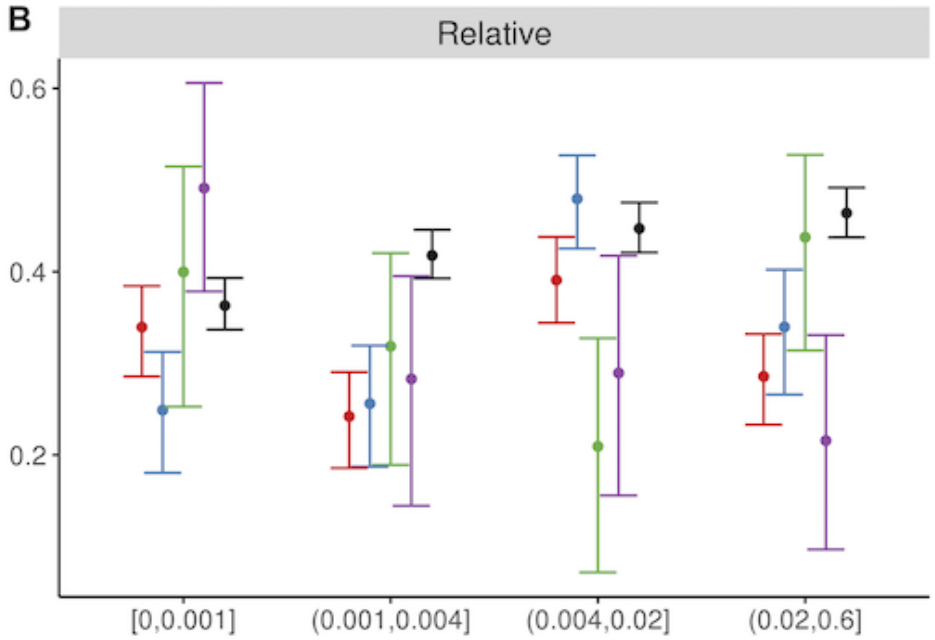
# Prediction 2: LOA in Africans is dependent of LD differences

$$Rel_{R2} = \frac{R^2_{bin}}{R^2_{total}}$$




Different slopes

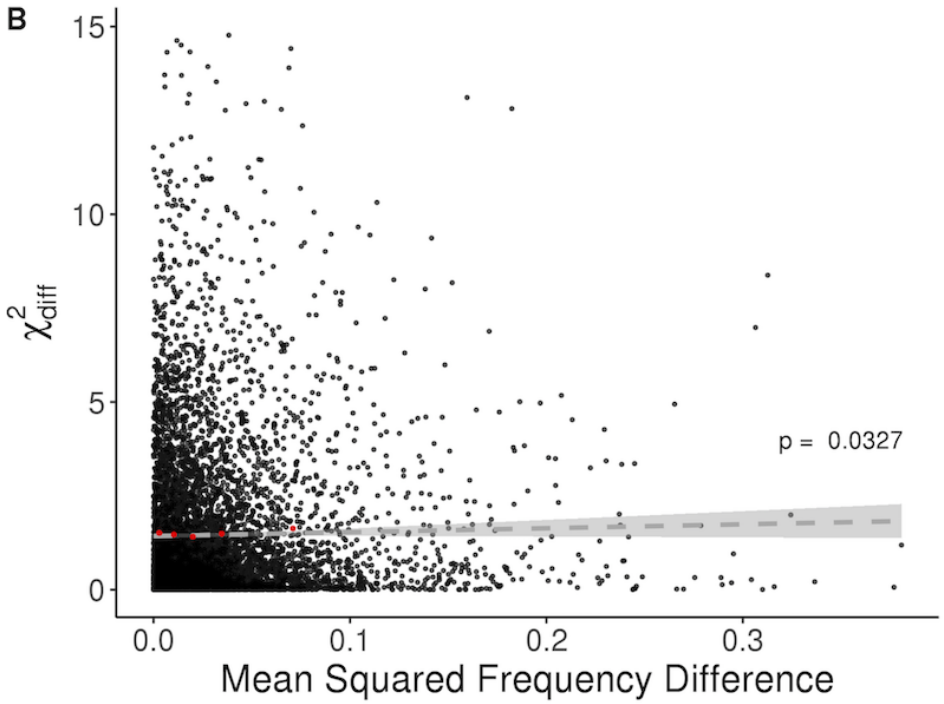



Recombination Rate

[Bitarello & Mathieson (2020), *G3*]

**Observation: loss of accuracy is somewhat dependent on recombination rate**

# Prediction: Differences in marginal effect sizes depend on allele frequency differences




[Bitarello & Mathieson (2020), *G3*]

$$N_{AFR} \sim 8,800$$

$$N_{EUR} \sim 350,000$$

$$\chi^2_{diff} = \left[ \frac{\beta_{eur} - \beta_{afr}}{\sqrt{SE^2_{eur} + SE^2_{afr}}} \right]^2$$

# Observation: Difference in marginal effect sizes increase with allele frequency differences across ancestries

# *Assuming* there are differences in marginal effect sizes

# *Assuming* there are differences in marginal effect sizes

$$PRS_1^C = \alpha PRS_{AFR} + (1-\alpha)PRS_{EUR}$$

Marquez-Luna *et al.* (2018) *Genet Epidem*

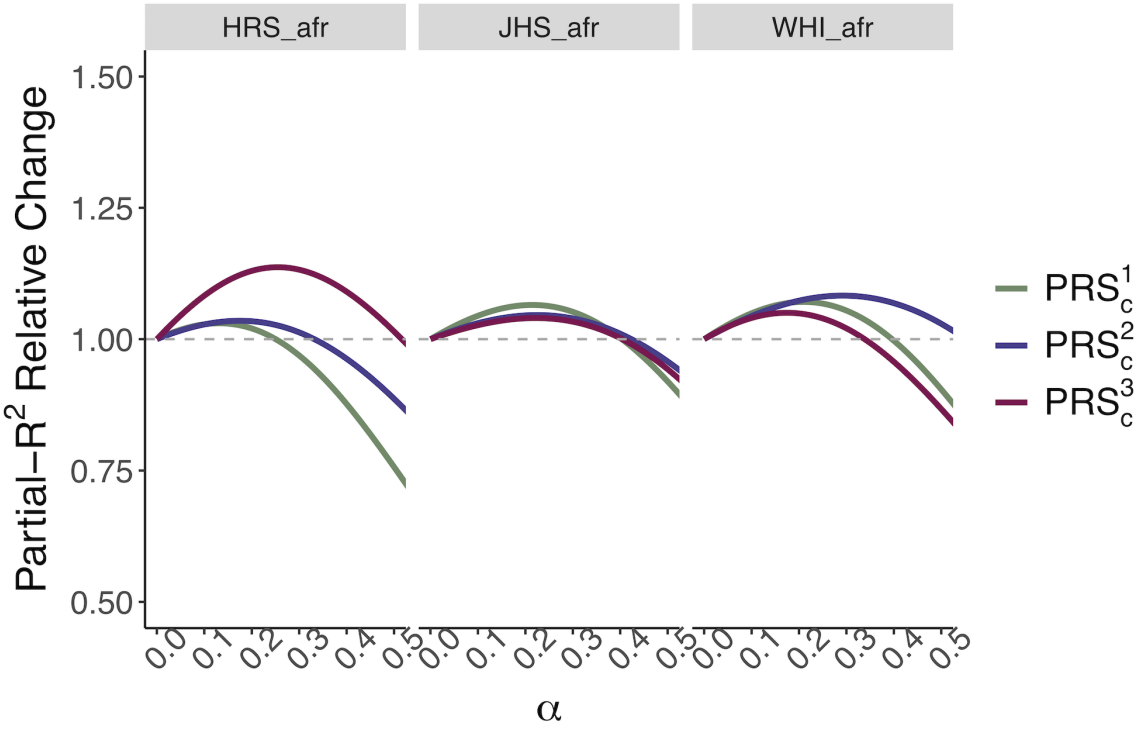
$$PRS_2^C = \alpha(1-p_{eur,j})PRS_{afr,j} + (1-\alpha+\alpha p_{eur,j})PRS_{EUR}$$

Bitarello & Mathieson (2020), *G3*

$$PRS_3^C = \alpha\left[\sum_{i\in AFR}\beta_{i,afr}G_i\right] + (1-\alpha)\left[\sum_{i\in AFR}\beta_{i,eur}G_i\right] + \left[\sum_{i\in EUR}\beta_{i,eur}G_i\right]$$

Bitarello & Mathieson (2020), *G3*

# Prediction: Including ancestry-specific effect sizes improves accuracy for admixed individuals




[Bitarello & Mathieson (2020), *G3*]

# Take Home Message

- LD and SFS impact PRS accuracy across ancestries
- not enough to explain loss of accuracy
- marginal effect sizes differences occur
- better testing with larger NEA sample sizes

## Future work

- optimizing ancestry-sensitive approach
- explore ancestry-dependent gene-gene interactions
- include more co-variates?

# Acknowledgements


Iain Mathieson

Arslan Zaidi

Dan Ju

Laura Colbran

Samantha Cox

Neale Lab

UK Biobank

Women's Health Initiative

Jackson Heart Study

Health and Retirement Study






# Thank you!

# Questions?