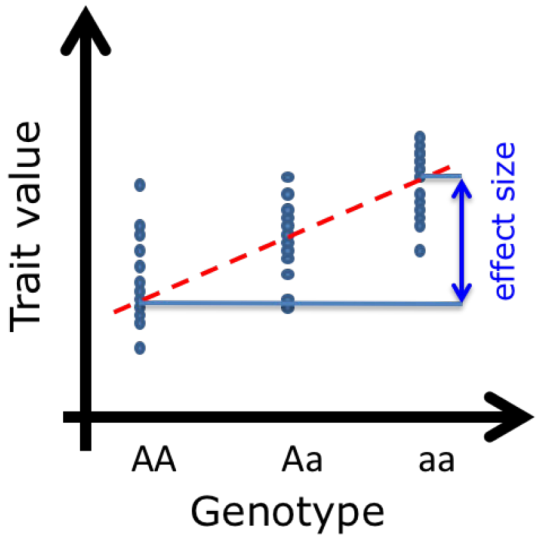
Beyond ancestry: the many factors influencing the portability of polygenic scores

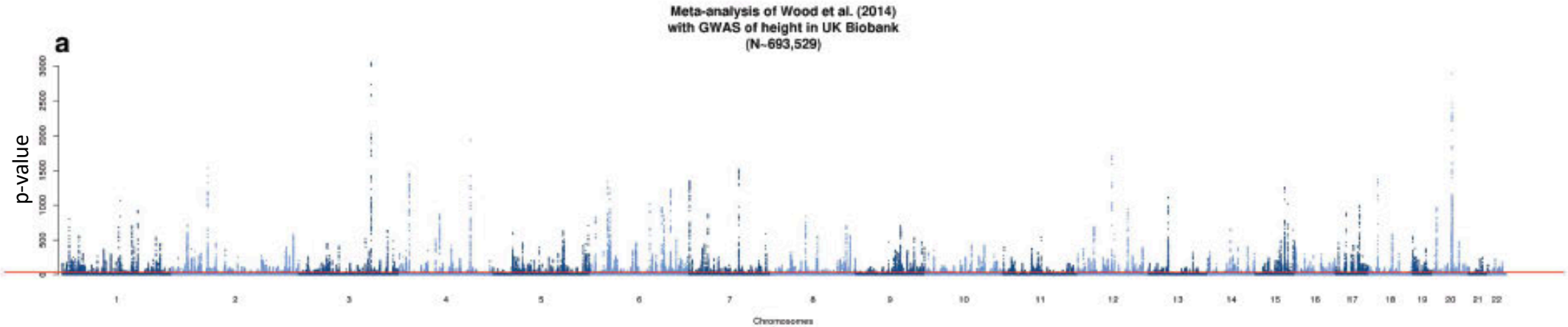
Molly Przeworski

Dept. of Biological Sciences
Dept. of Systems Biology
Columbia University

In three parts

- Why genetic ancestry impacts prediction accuracies
- Prediction accuracy can vary, a lot, even within an ancestry
- GWAS signals need to be deconstructed to understand what is going on


Trait value
effect size
AA
Aa
aa
Genotype


Meta-analysis of Wood et al. (2014)
with GWAS of height in UK Biobank
(N~693,529)
a
p-value
Chromosomes

Polygenic scores

genotype ∈ {0,1,2}

$$\sum_{\{sites\ i\}} X_i \hat{\beta}_i$$

estimated effect

Polygenic scores

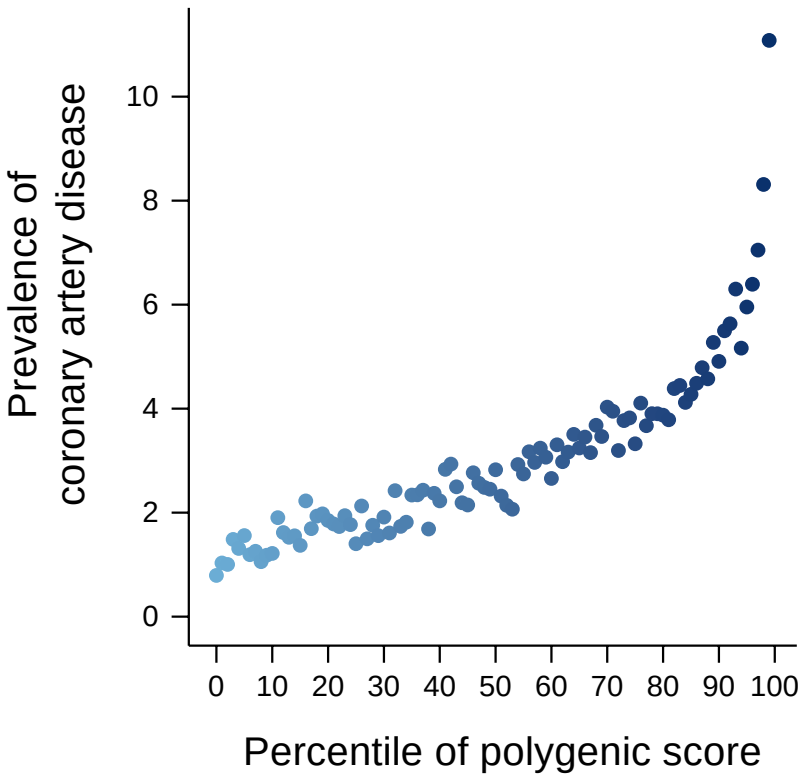
genotype ∈ {0,1,2}

$$\sum_{\{sites\ i\}} X_i \hat{\beta}_i \longrightarrow \text{trait value}$$

Prediction accuracy :

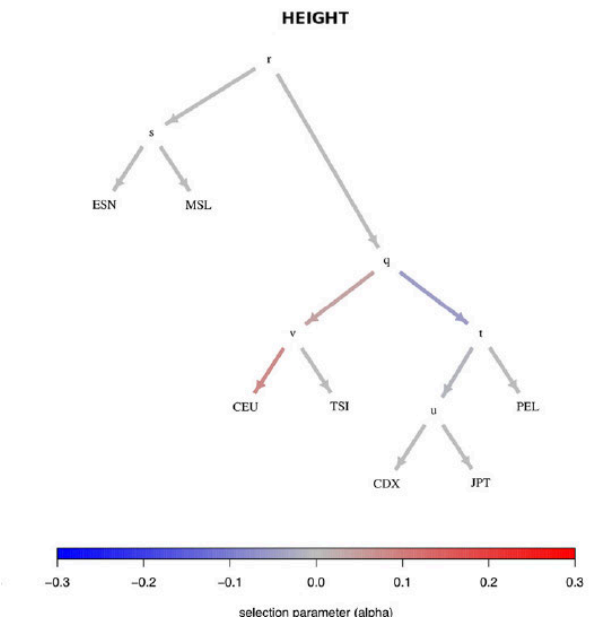
Corr (PGS, trait value)

In human genetics



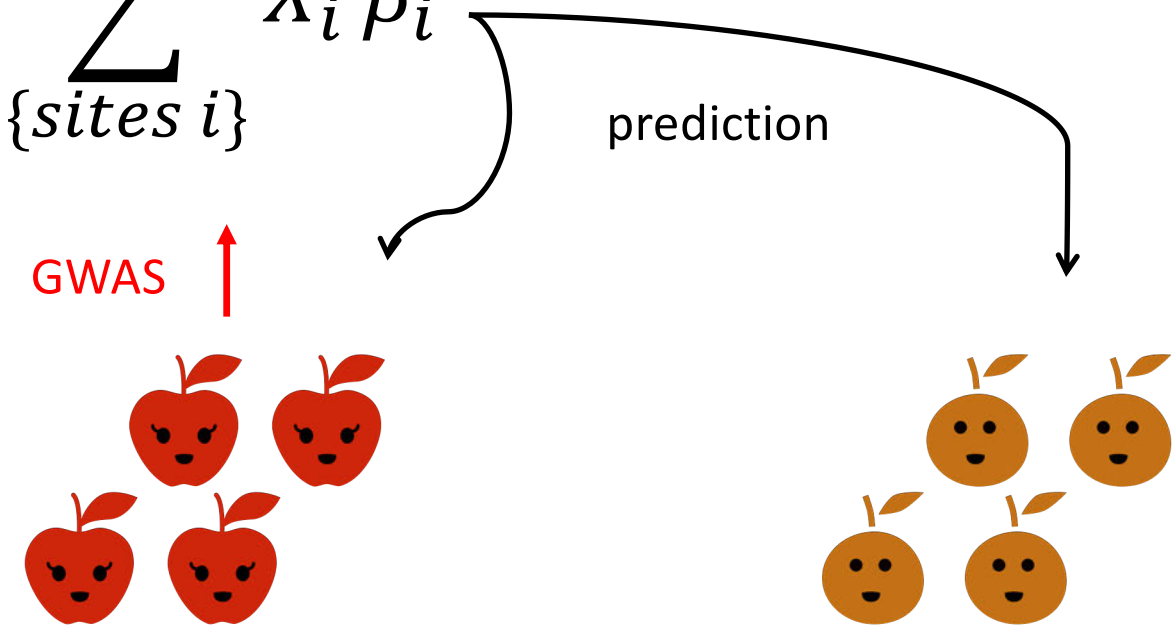
From Khera et al. 2018

In population genetics



From Racimo et al. 2018

$$\sum_{\{sites\ i\}} X_i \hat{\beta}_i$$



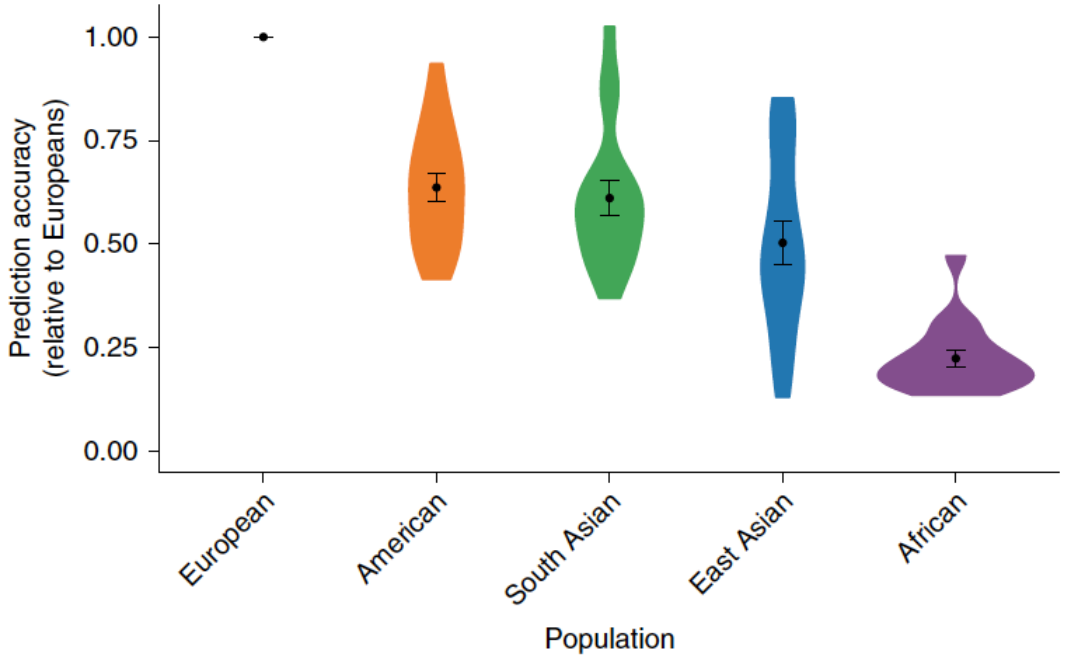


Fig. 3 | Prediction accuracy relative to European-ancestry individuals across 17 quantitative traits and 5 continental populations in the UKBB.

Martin et al., 2019

What is going on?

<u>Population genetic explanations</u>

Linkage disequilibrium patterns differ among populations

Allele frequencies change over time

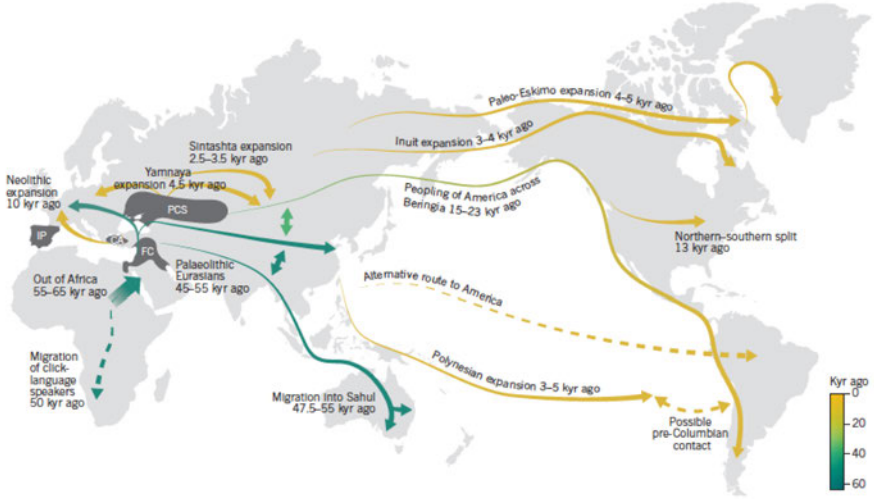


Figure 3 | Major human migrations across the world inferred through analyses of genomic data. Some migration routes remain under debate. For example, there is still some uncertainty regarding the migration routes used to populate the Americas. Genomic data are limited in their resolution to determine paths of migration because further population movements, subsequent to the initial migrations, may obscure the geographic patterns that can be discerned from the genomic data. Proposed routes of migration that remain controversial are indicated by dashed lines. CA, Central Anatolia; FC, Fertile Crescent; IP, Iberian Peninsula; PCS, Pontic–Caspian steppe.

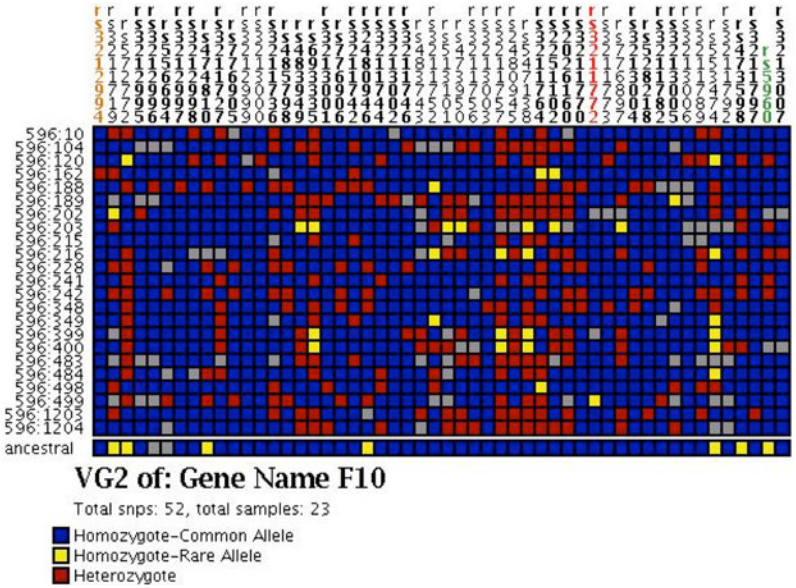
Nielsen et al. 2017

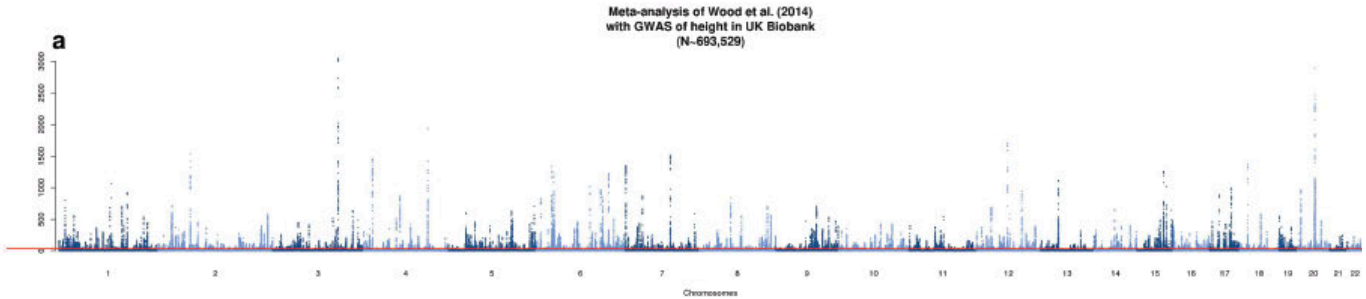
What is going on?

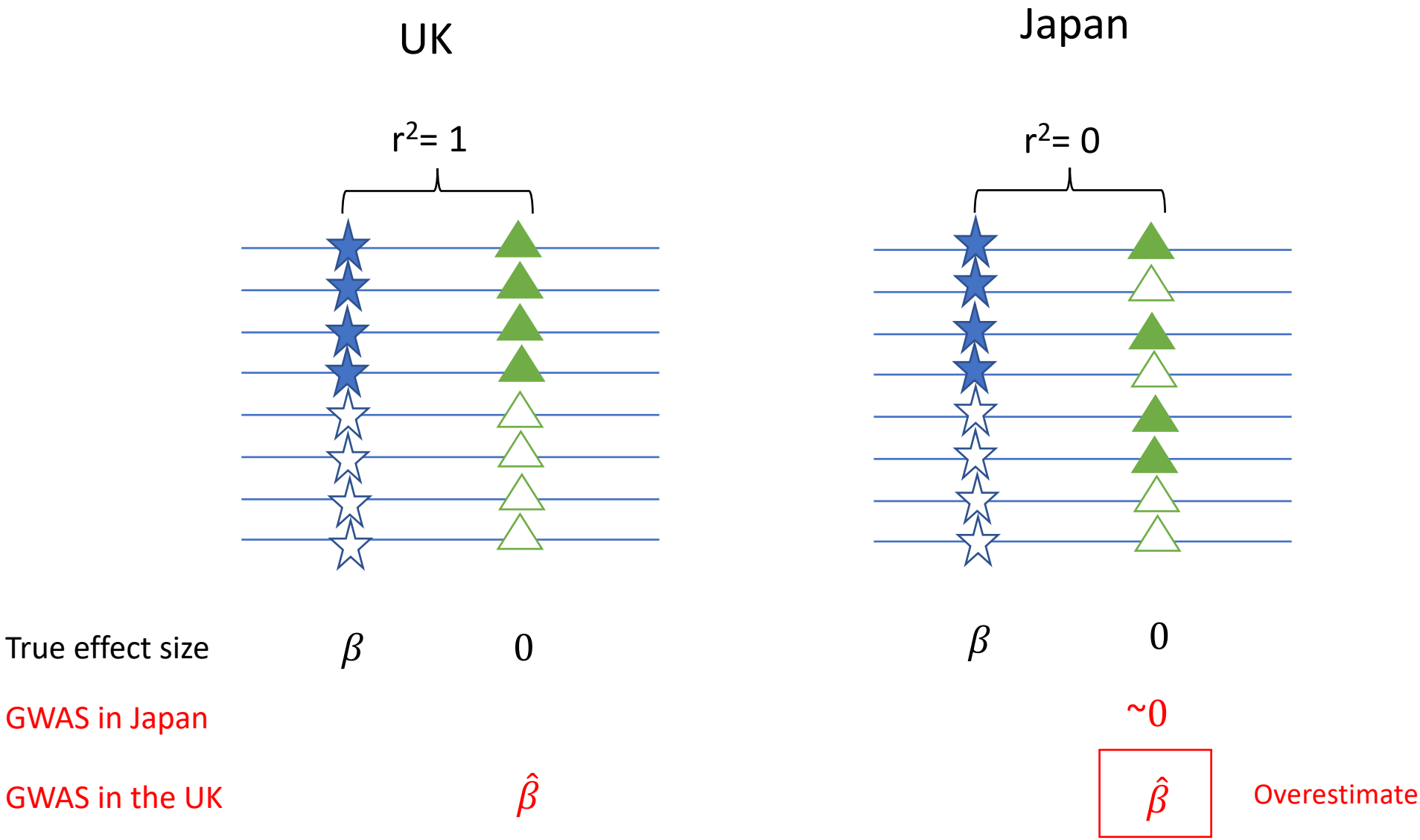
<u>Population genetic explanations</u>

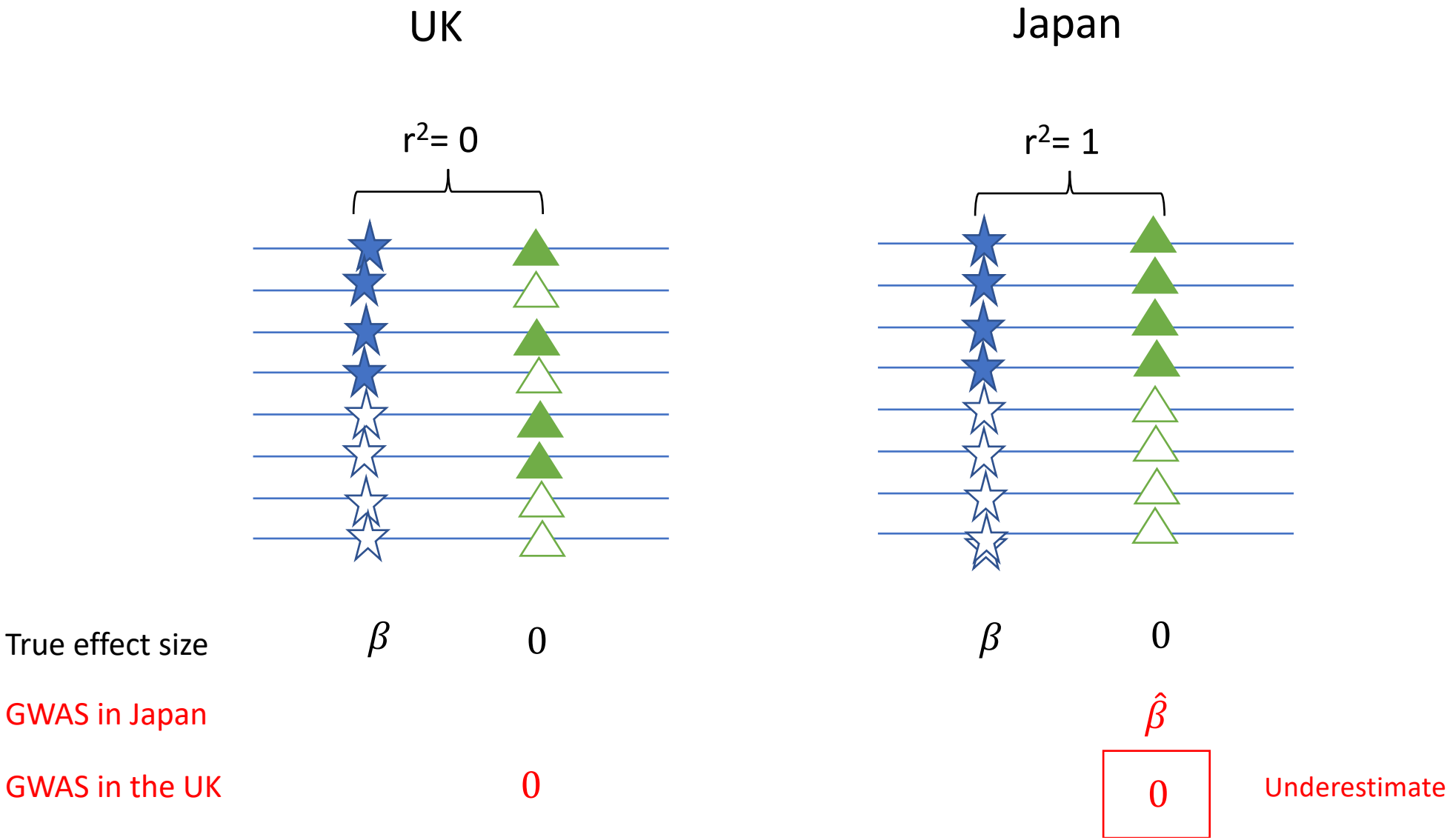
Linkage disequilibrium patterns differ among populations

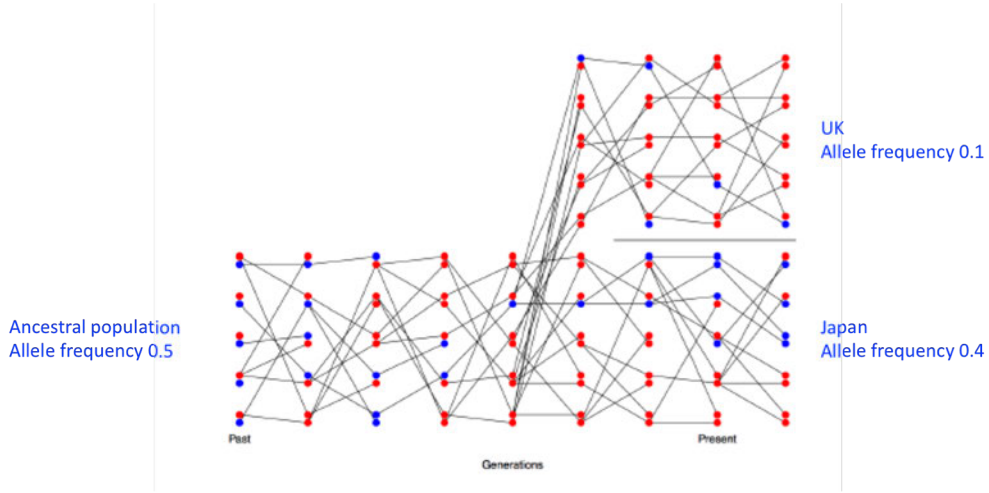
Allele frequencies change over time






UK
Japan
r2= 1
r2= 0
True effect size
β
0
β
0
GWAS in Japan
~0
GWAS in the UK
β̂
β̂
Overestimate


UK
Japan
r2= 0
r2= 1
True effect size
β
0
β
0
GWAS in Japan
β̂
GWAS in the UK
0
0
Underestimate



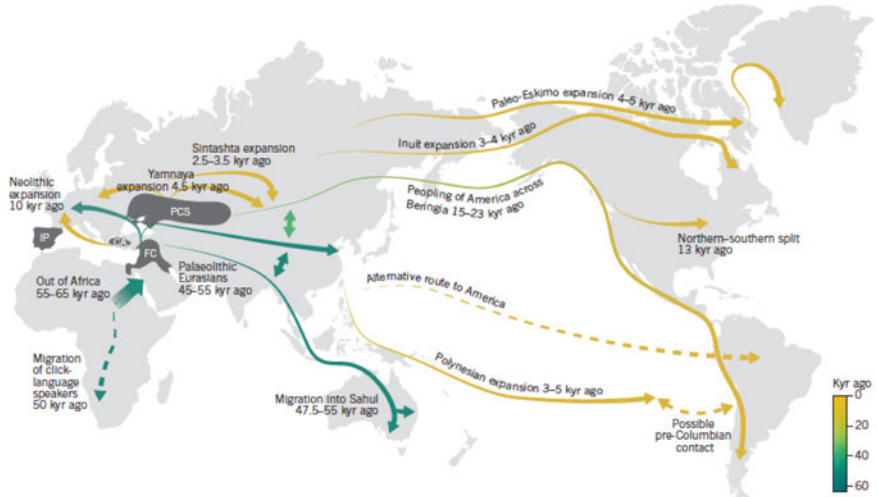


Figure 3 | Major human migrations across the world inferred through analyses of genomic data. Some migration routes remain under debate. For example, there is still some uncertainty regarding the migration routes used to populate the Americas. Genomic data are limited in their resolution to determine paths of migration because further population movements, subsequent to the initial migrations, may obscure the geographic patterns that can be discerned from the genomic data. Proposed routes of migration that remain controversial are indicated by dashed lines. CA, Central Anatolia; FC, Fertile Crescent; IP, Iberian Peninsula; PCS, Pontic–Caspian steppe.

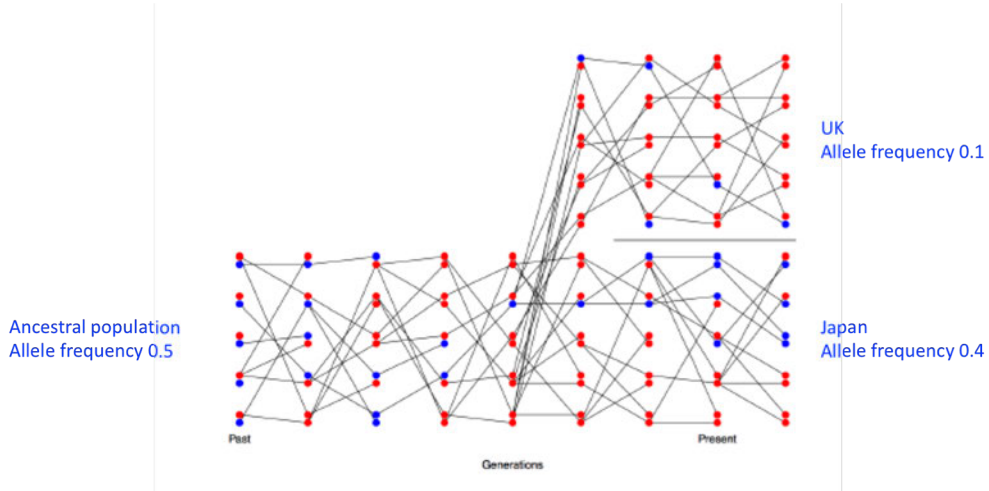
Nielsen et al. 2017

$$S_i = \sum \hat{\beta}_j g_{ij}$$

Polygenic score of individual *i*

Genotype of individual *i* at locus *j*

Effect size of genetic variant *j* on the phenotype



$$S_i = \sum \hat{\beta}_j g_{ij}$$

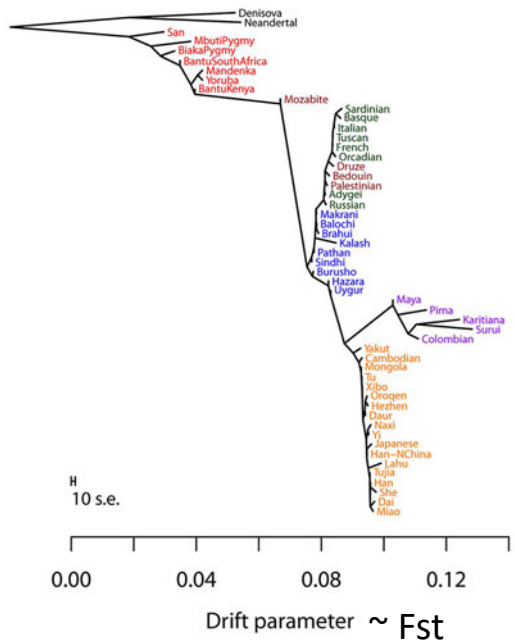
Polygenic score of individual i

Genotype of individual i at locus j

Effect size of genetic variant j on the phenotype

⇒ Polygenic scores will diverge by genetic drift alone

Genetic variance explained decreases with Fst



Pickrell & Pritchard 2012

⇒ Polygenic score will also diverge if the trait (or a correlated trait) is under selection

Population genetic explanations

Linkage disequilibrium patterns differ among populations

Allele frequencies change over time

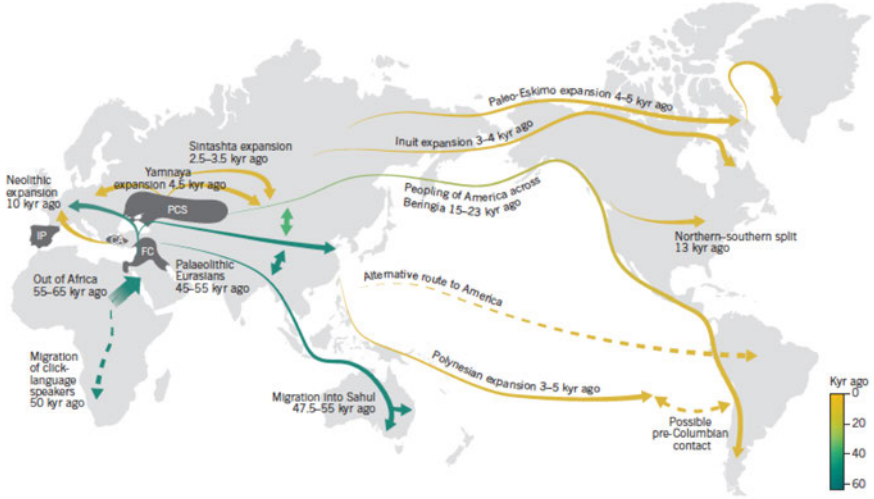
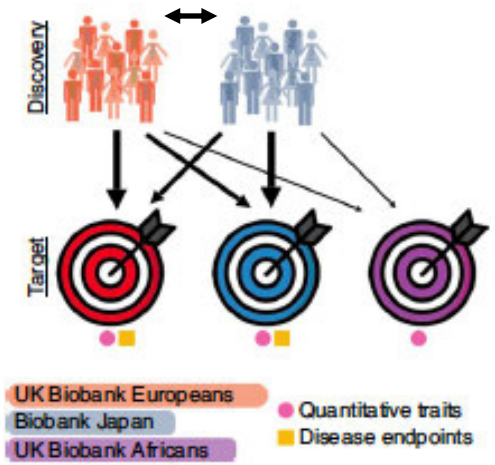


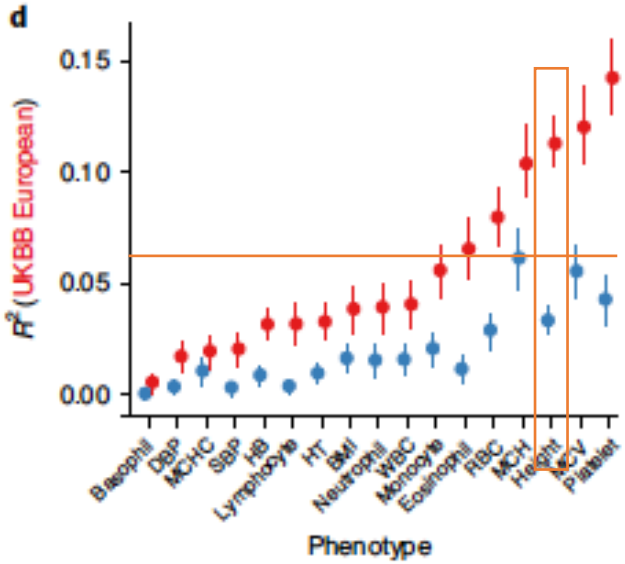
Figure 3 | Major human migrations across the world inferred through analyses of genomic data. Some migration routes remain under debate. For example, there is still some uncertainty regarding the migration routes used to populate the Americas. Genomic data are limited in their resolution to determine paths of migration because further population movements, subsequent to the initial migrations, may obscure the geographic patterns that can be discerned from the genomic data. Proposed routes of migration that remain controversial are indicated by dashed lines. CA, Central Anatolia; FC, Fertile Crescent; IP, Iberian Peninsula; PCS, Pontic–Caspian steppe.

Nielsen et al. 2017

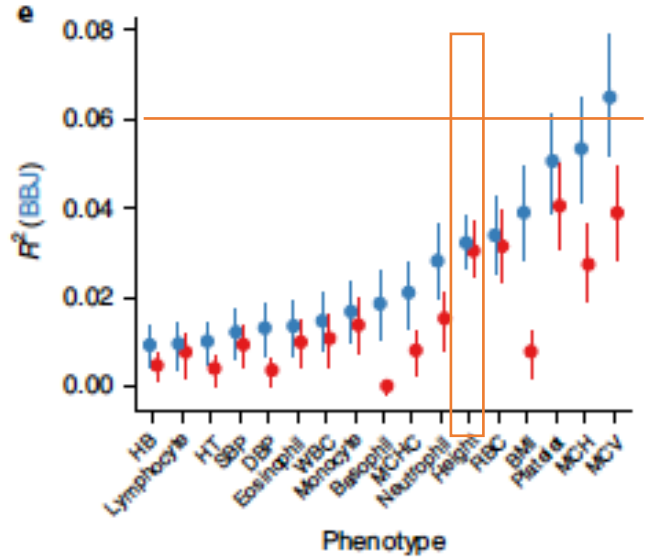
matched sample sizes



GWAS in the UK



GWAS in Japan



Martin et al., 2019

What is going on?

~~Population genetic explanations~~

~~Linkage disequilibrium patterns differ~~

~~Allele frequency changes~~

Other possibilities

Differences in *Ve*

$$y = \hat{\beta}x + \epsilon$$

Gene by environment interactions

~~Population stratification~~

See Berg et al. 2019; Sohail et al. 2019; Nordborg et al. 2019

Hakhamanesh Mostafavi

Arbel Harpak

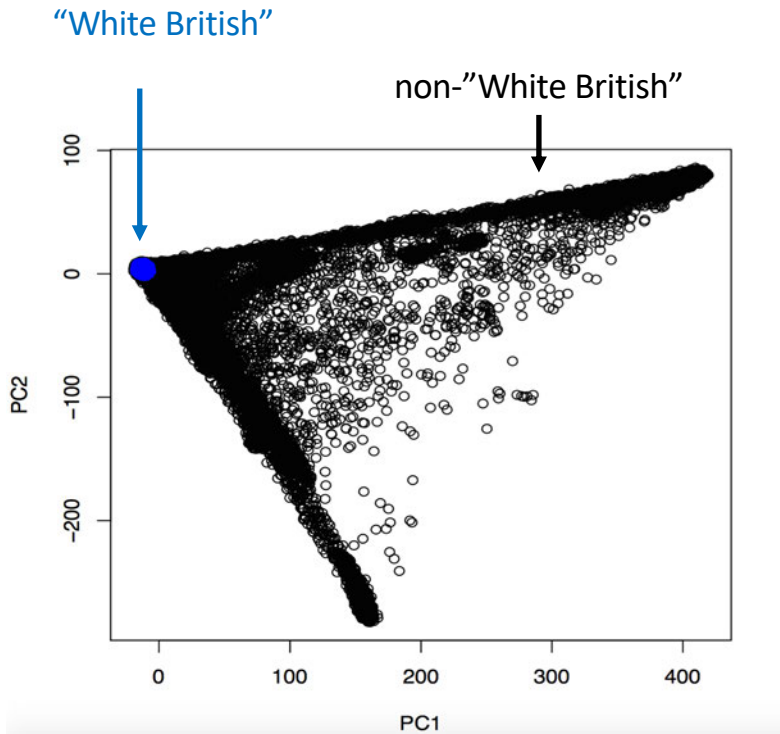
Ipsita Agarwal

Dalton Conley

Jonathan Pritchard

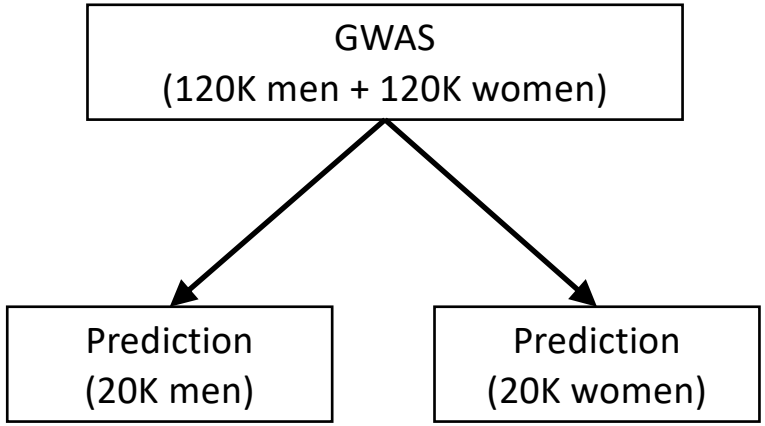
Using the UK Biobank

- 340K unrelated “White British” individuals
- 20K sibling pairs



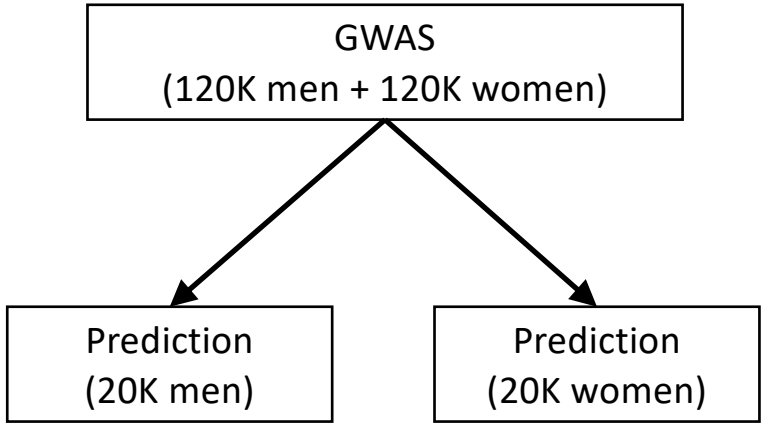
Example 1: prediction accuracy of **blood pressure** by **sex**

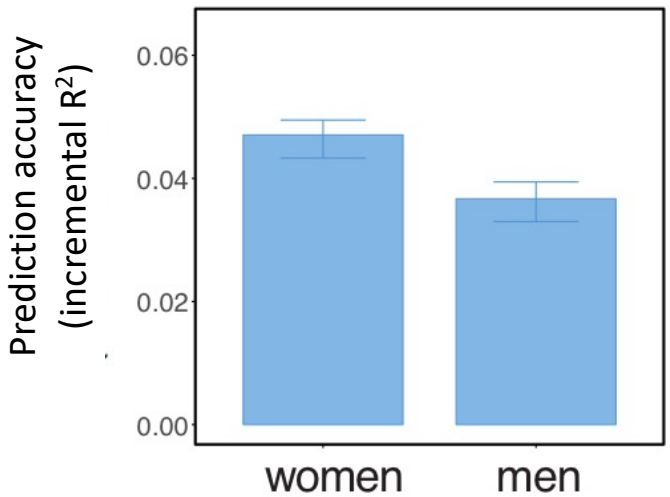
Diastolic blood pressure



Example 1: prediction accuracy of **blood pressure** by **sex**

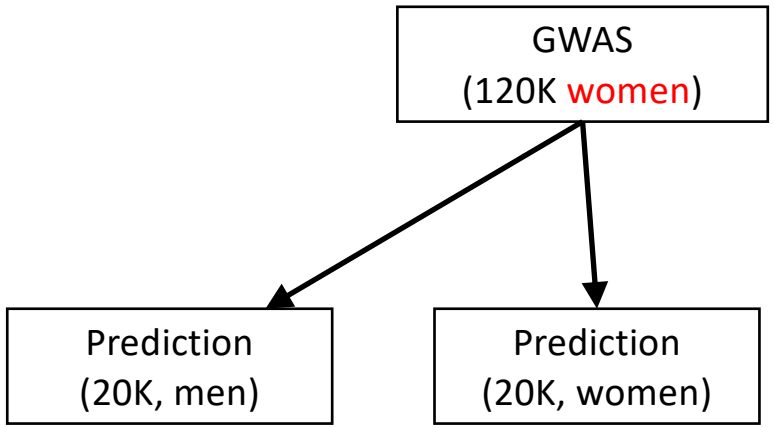
Diastolic blood pressure

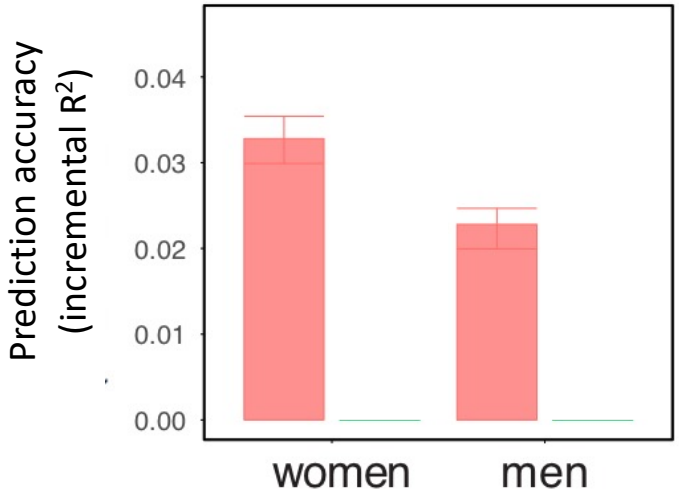




Prediction accuracy depends on characteristics of both GWAS and prediction set

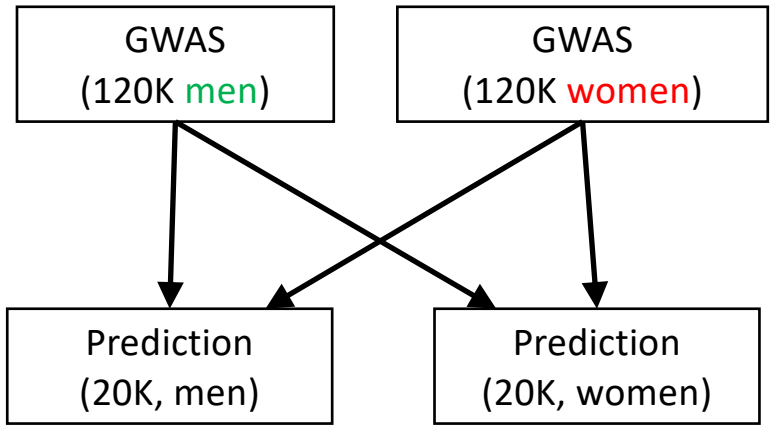
Diastolic blood pressure

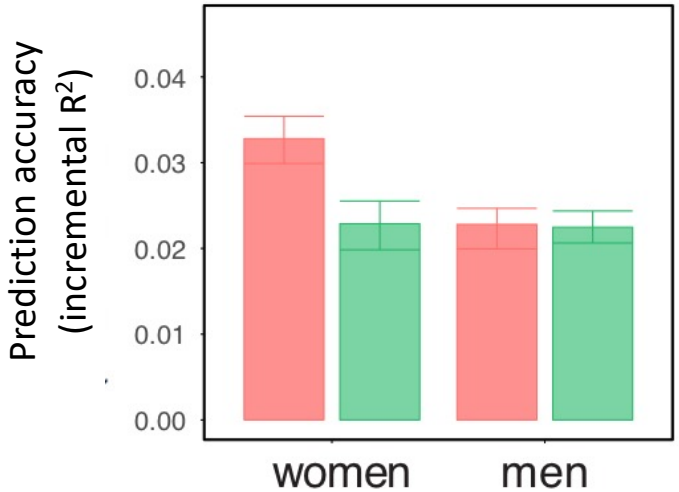




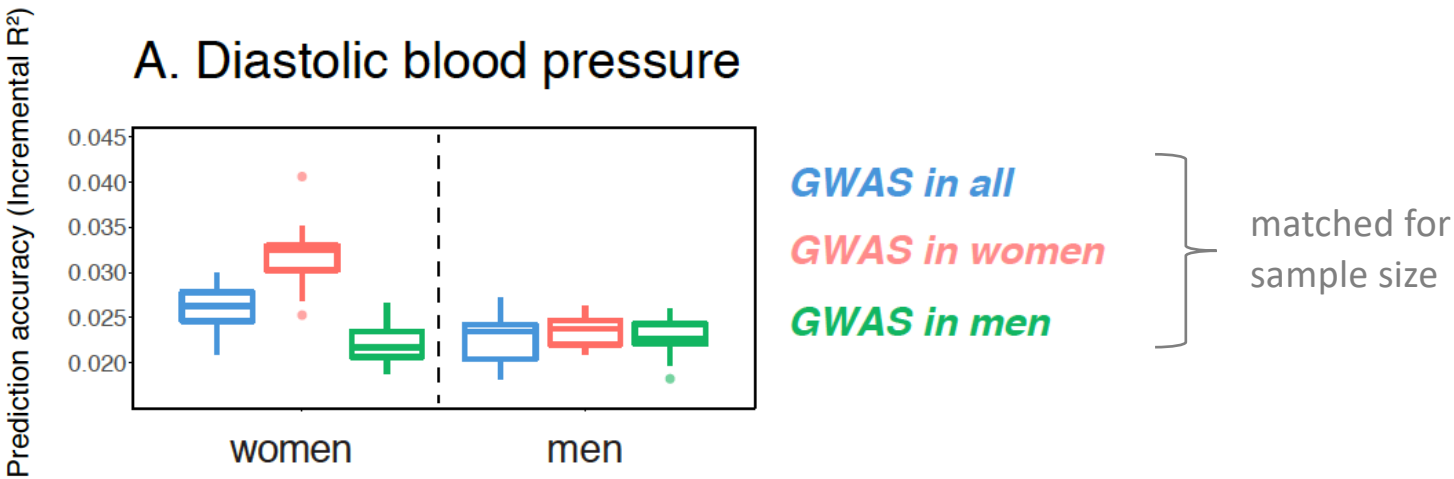
Prediction accuracy depends on characteristics of both GWAS and prediction set

Diastolic blood pressure

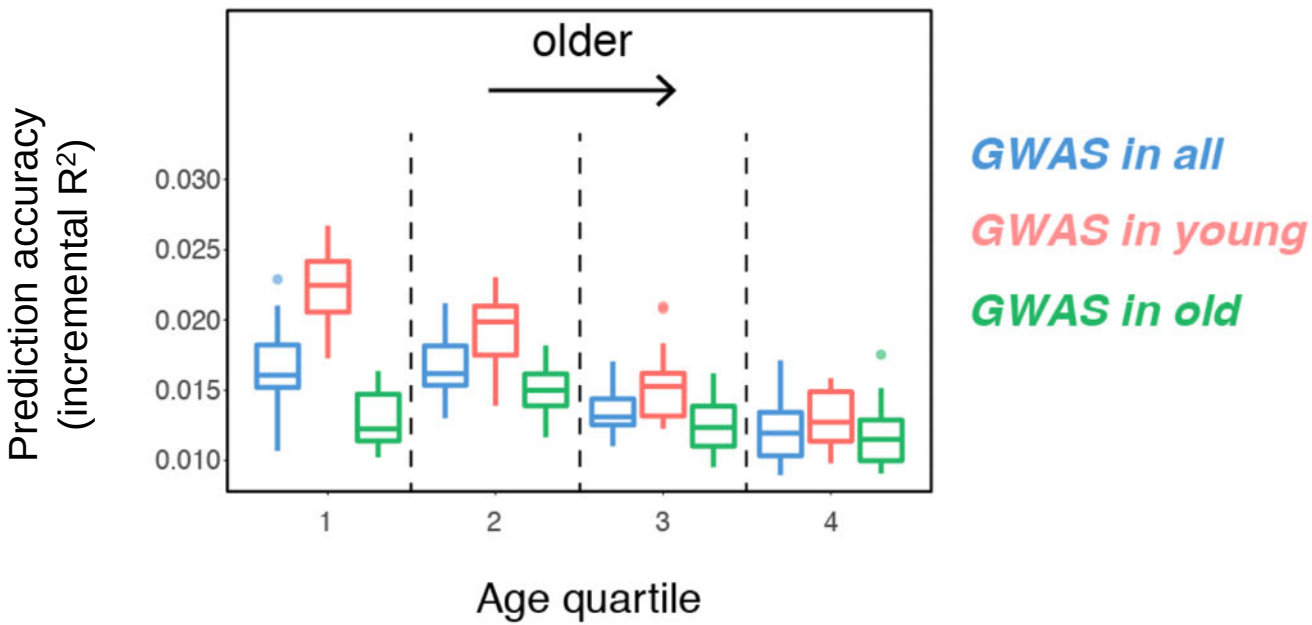




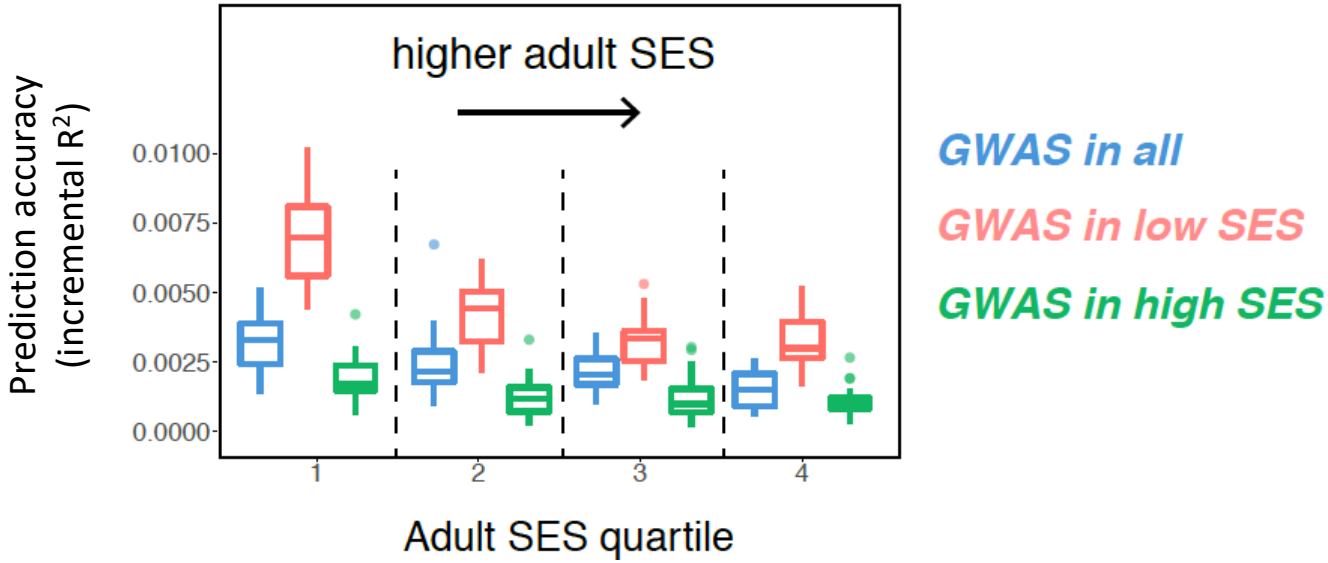
Prediction accuracy depends on characteristics of both GWAS and prediction set



Example 2: Prediction accuracy for **BMI** varies by **age group**

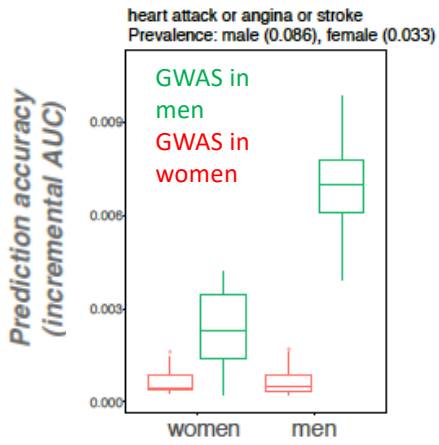


Example 3: Prediction accuracy for **years of schooling** varies by **socioeconomic status (SES)**



Robust to various sensitivity analyses:

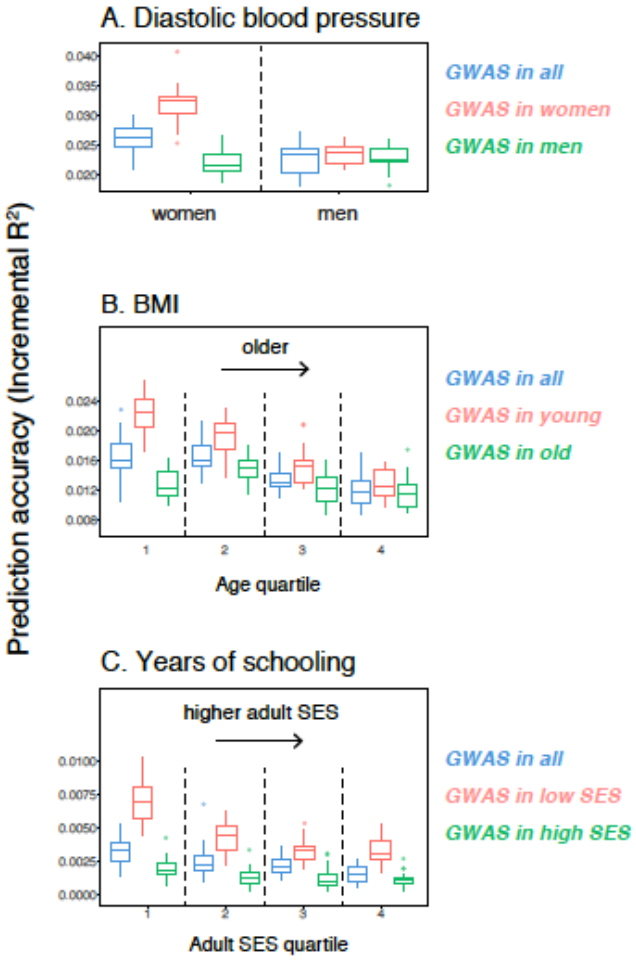
- method and parameters used for **GWAS**
- method and parameters to build **polygenic score**
- Metric to evaluate **prediction accuracy**
- **Disease/binary phenotypes;** not just continuous phenotypes



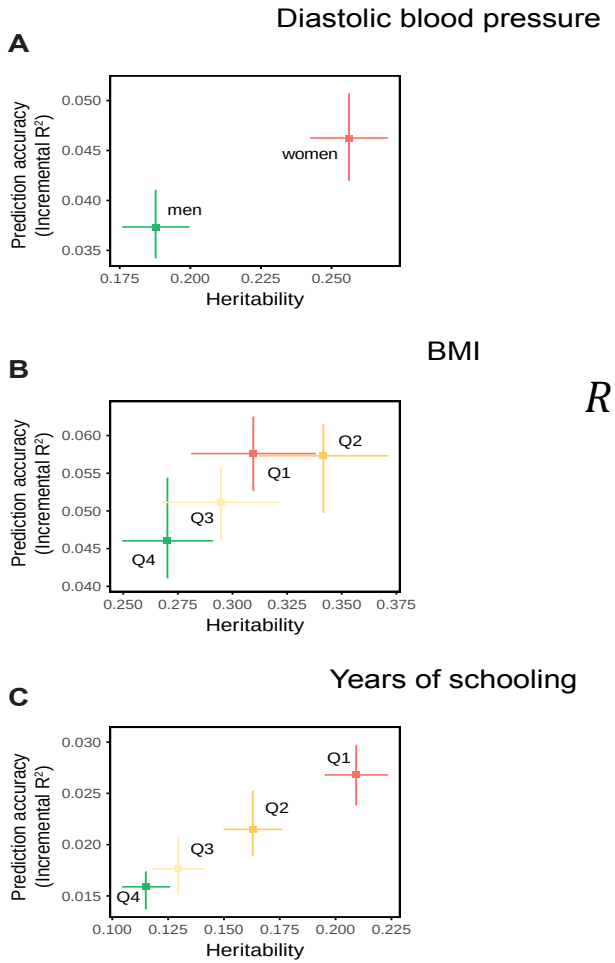
Prediction accuracy depends on sample characteristics

For many traits,
we do not know what sample characteristics matter.

What is going on in these examples?

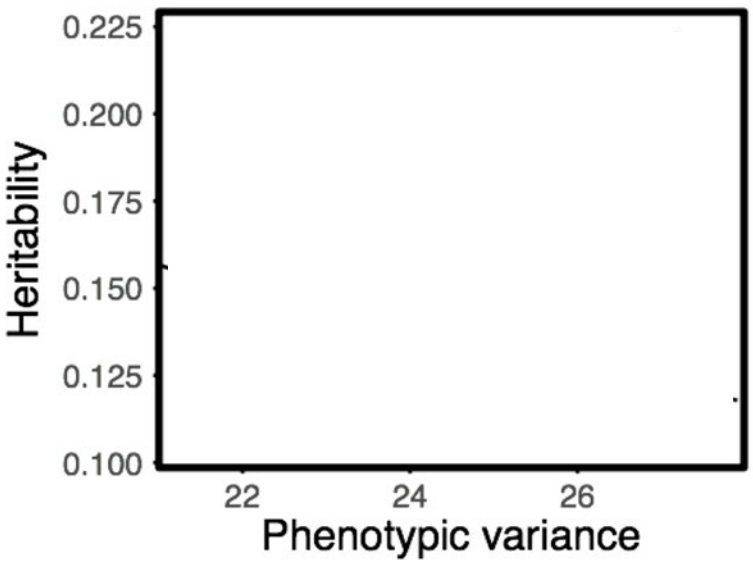


Prediction accuracies track (SNP) heritabilities



$$R^2 \propto \frac{Vg}{Vp} = \frac{Vg}{Vg + Ve}$$

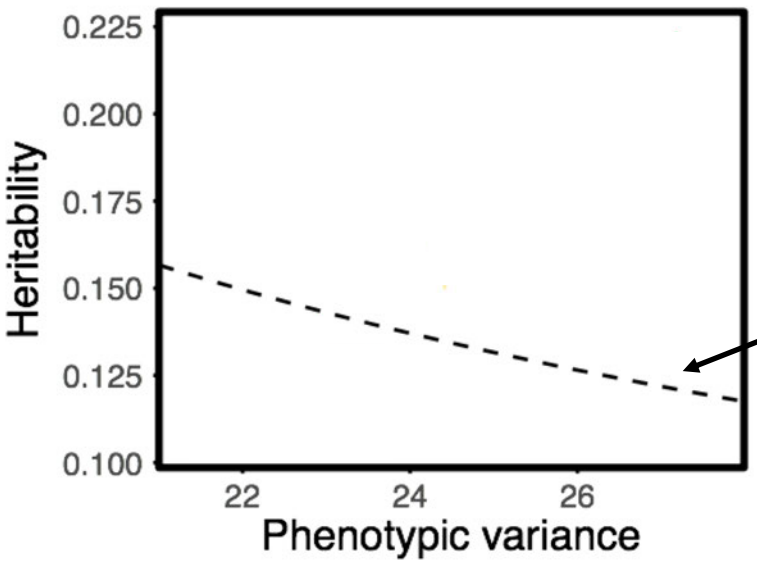
Simplest possibility: heritabilities vary across strata because the environmental variance does



$$R^2 \propto \frac{Vg}{Vp} = \frac{Vg}{Vg + Ve}$$

Are the heritabilities across strata reflecting different environmental variances?

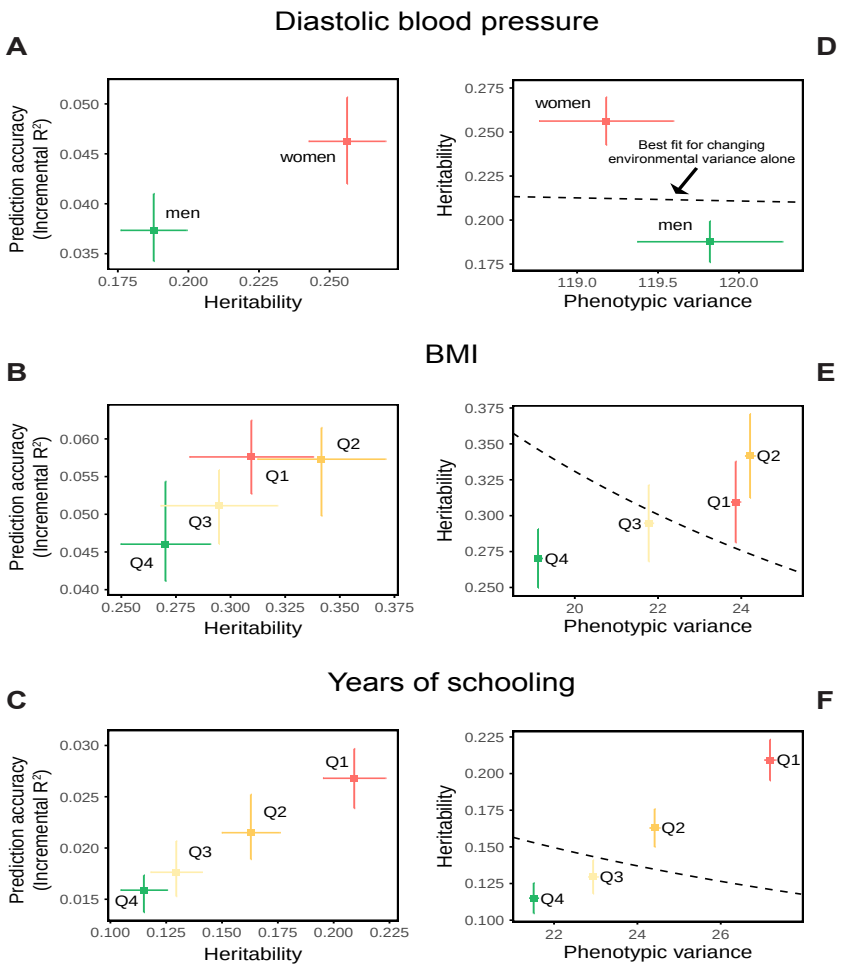
$$R^2 \propto \frac{Vg}{Vp} = \frac{Vg}{Vg + \boxed{Ve}}$$



Expectation under changes in *Ve* alone

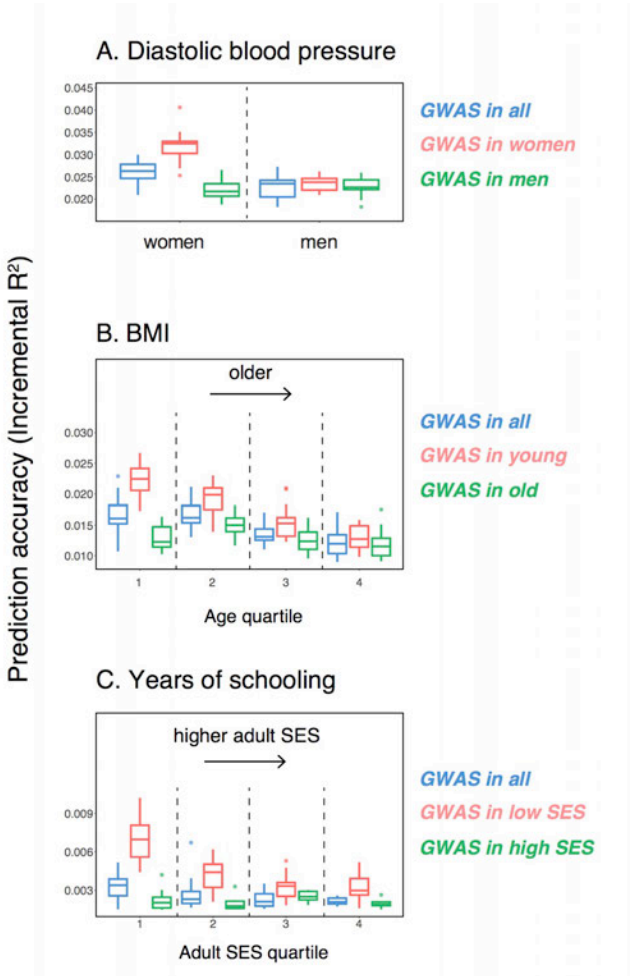
Greater phenotypic variance => lower heritability

Not just a difference in environmental variances

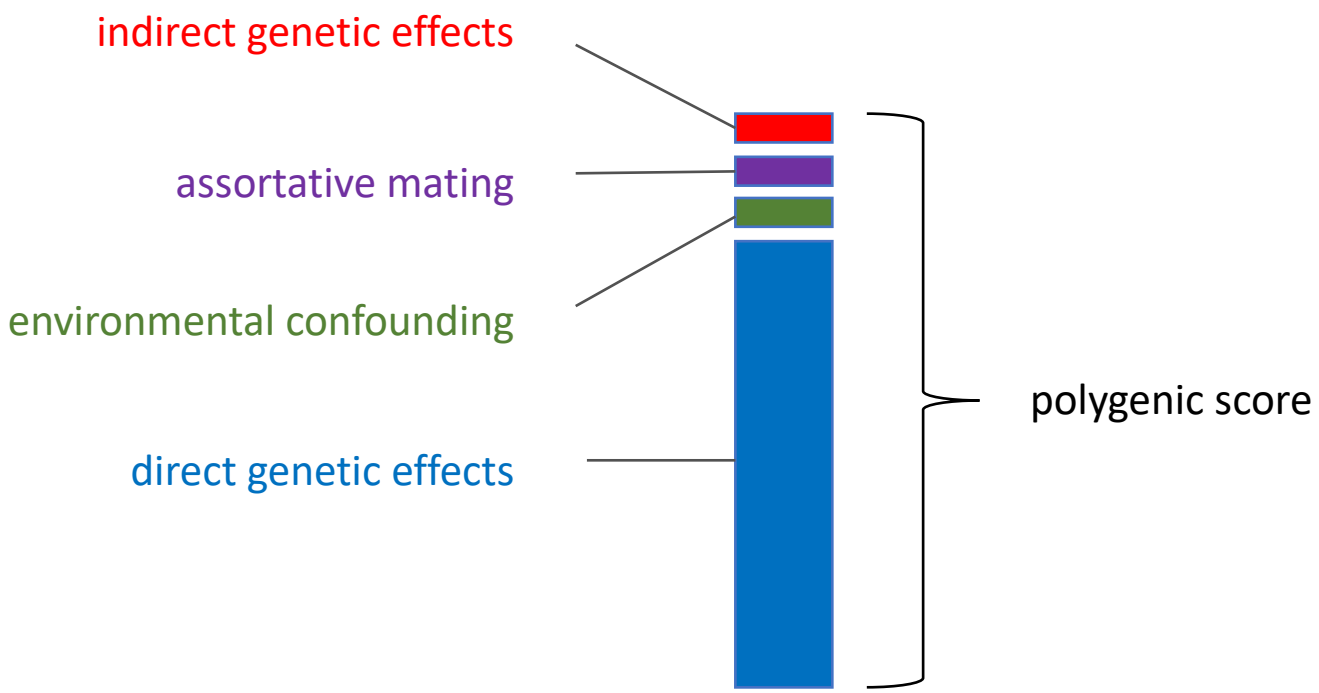


- Genetic effects highly correlated across strata
- Genetic amplification in some strata?

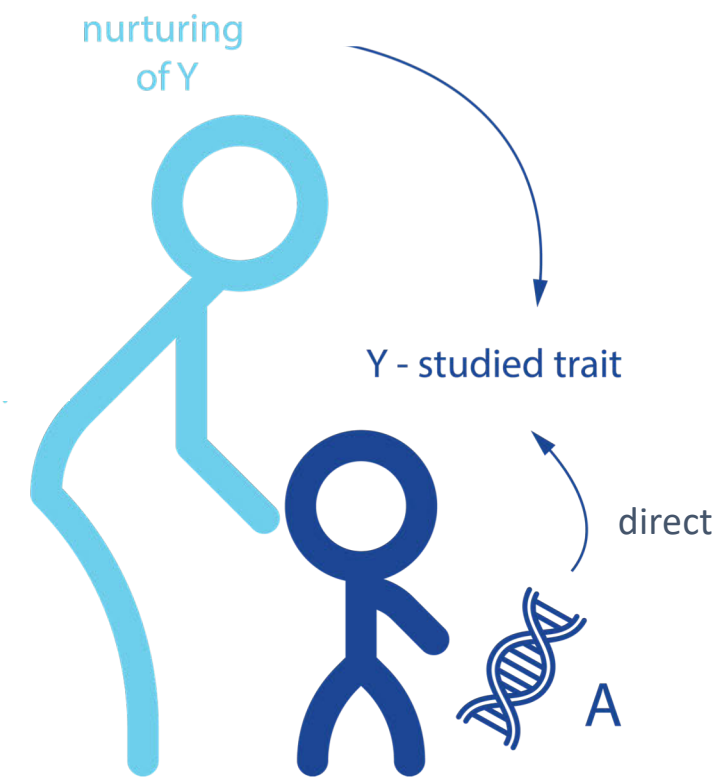
What is going on in these examples?



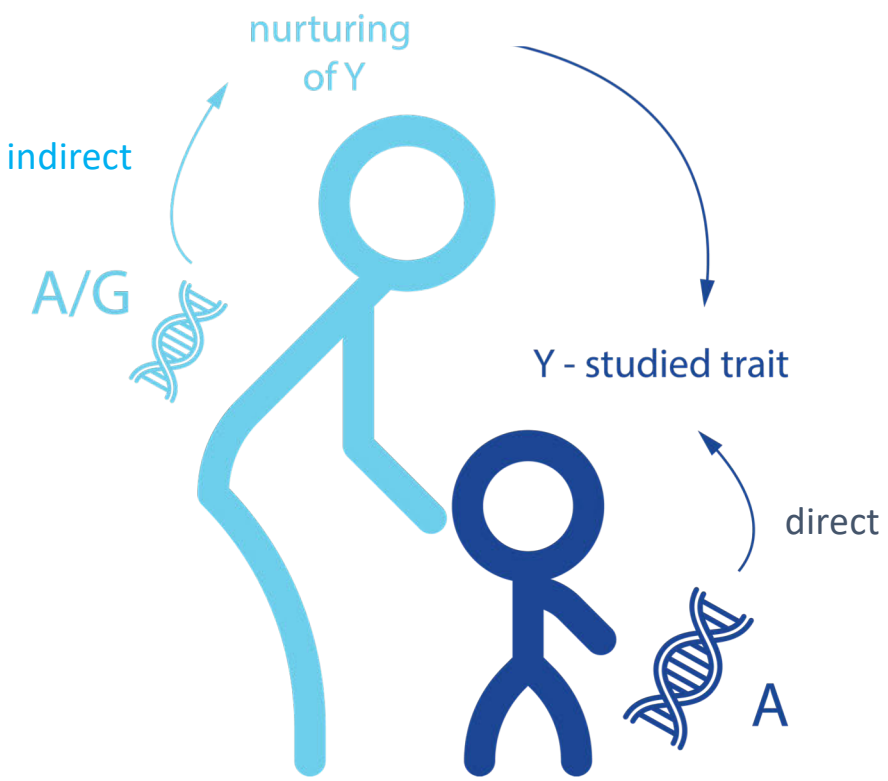
GWAS pick up more than just direct genetic effects



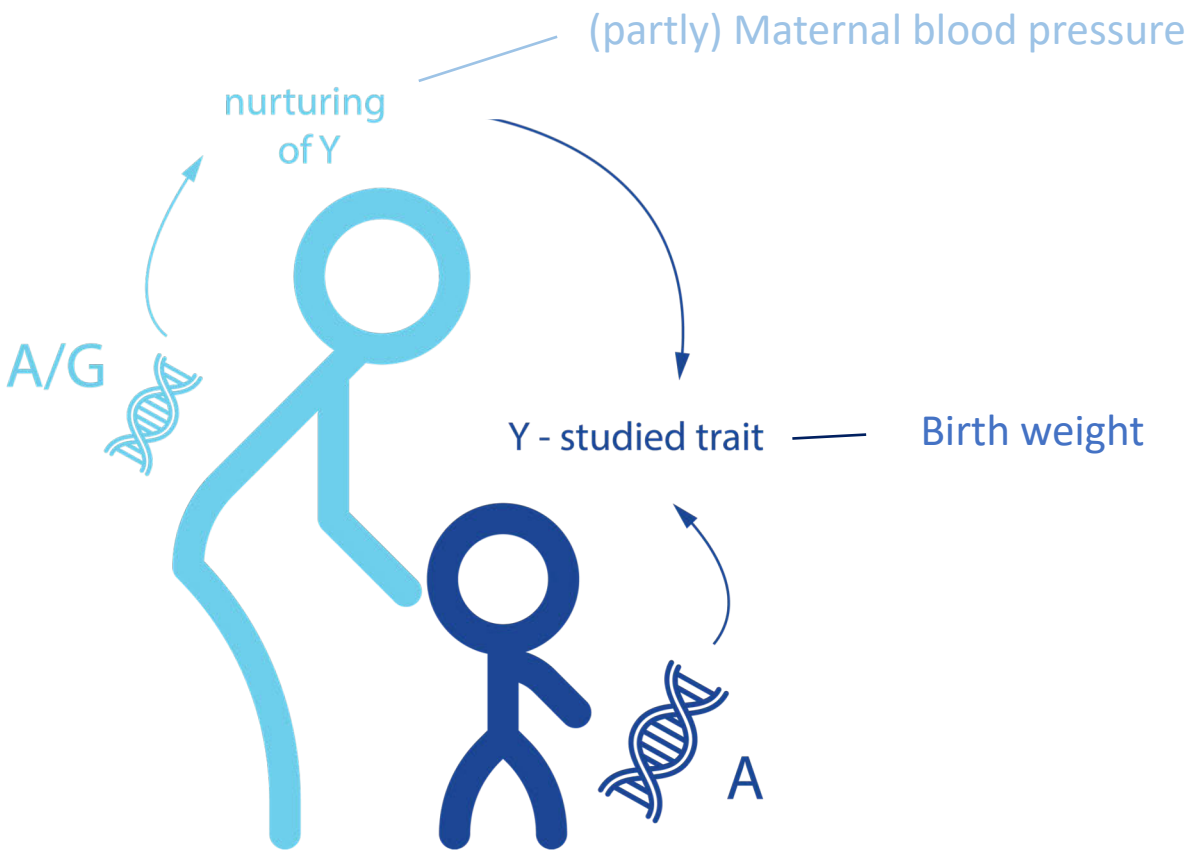
Example: GWAS also pick up indirect genetic effects



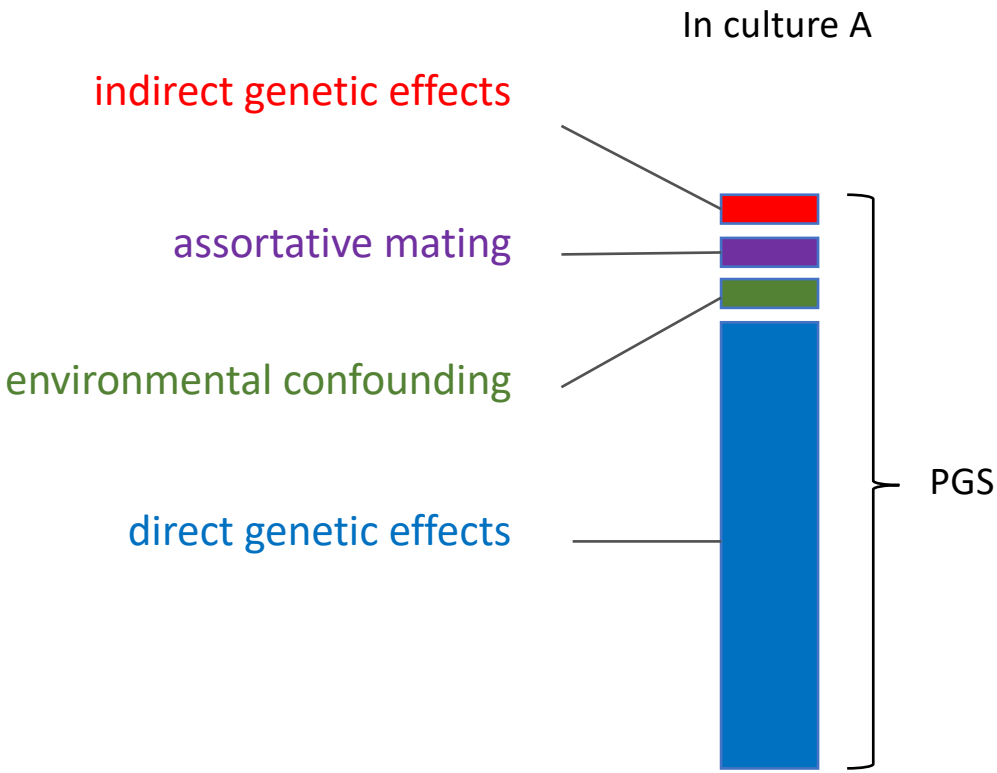
Example: GWAS also pick up indirect genetic effects

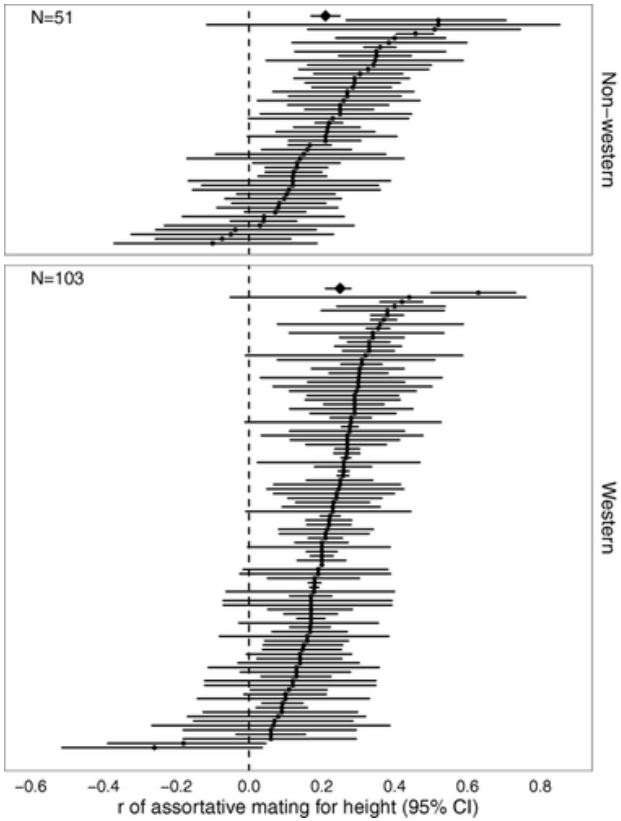


Example: GWAS also pick up indirect genetic effects



Warrington et al. 2019





Stulp et al. 2016

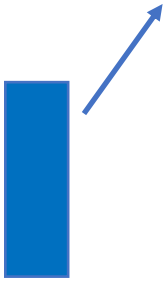
Do these other effects contribute?

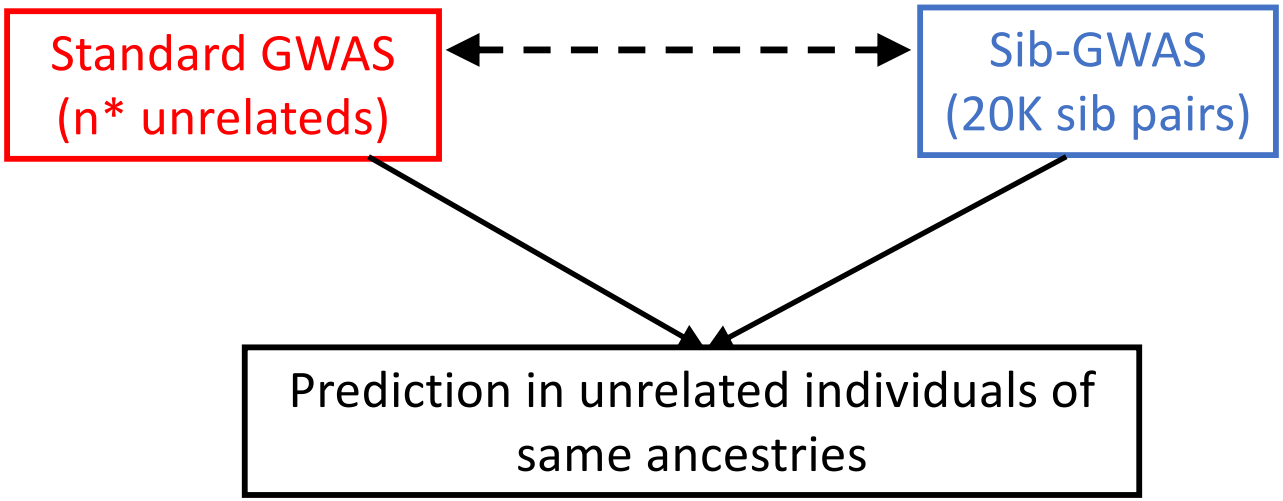
phenotype

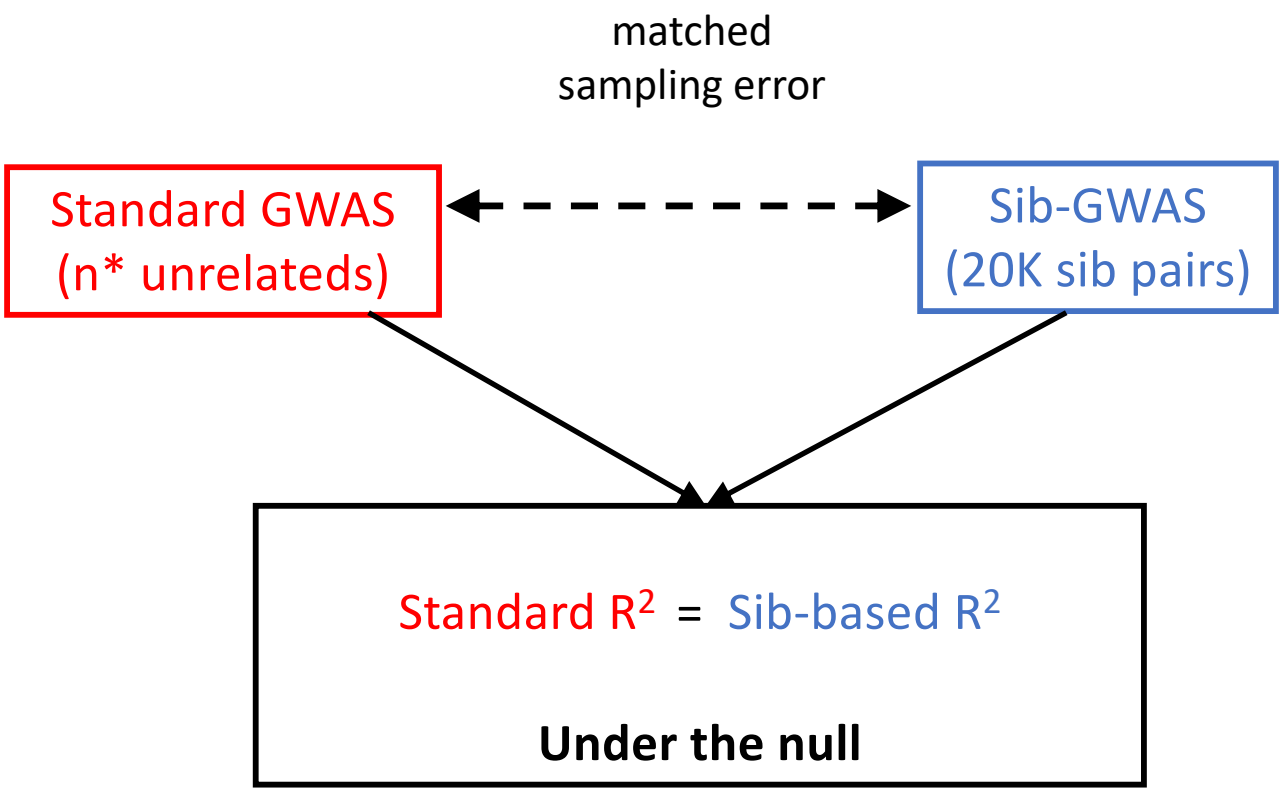
genotype ∈ {0,1,2}

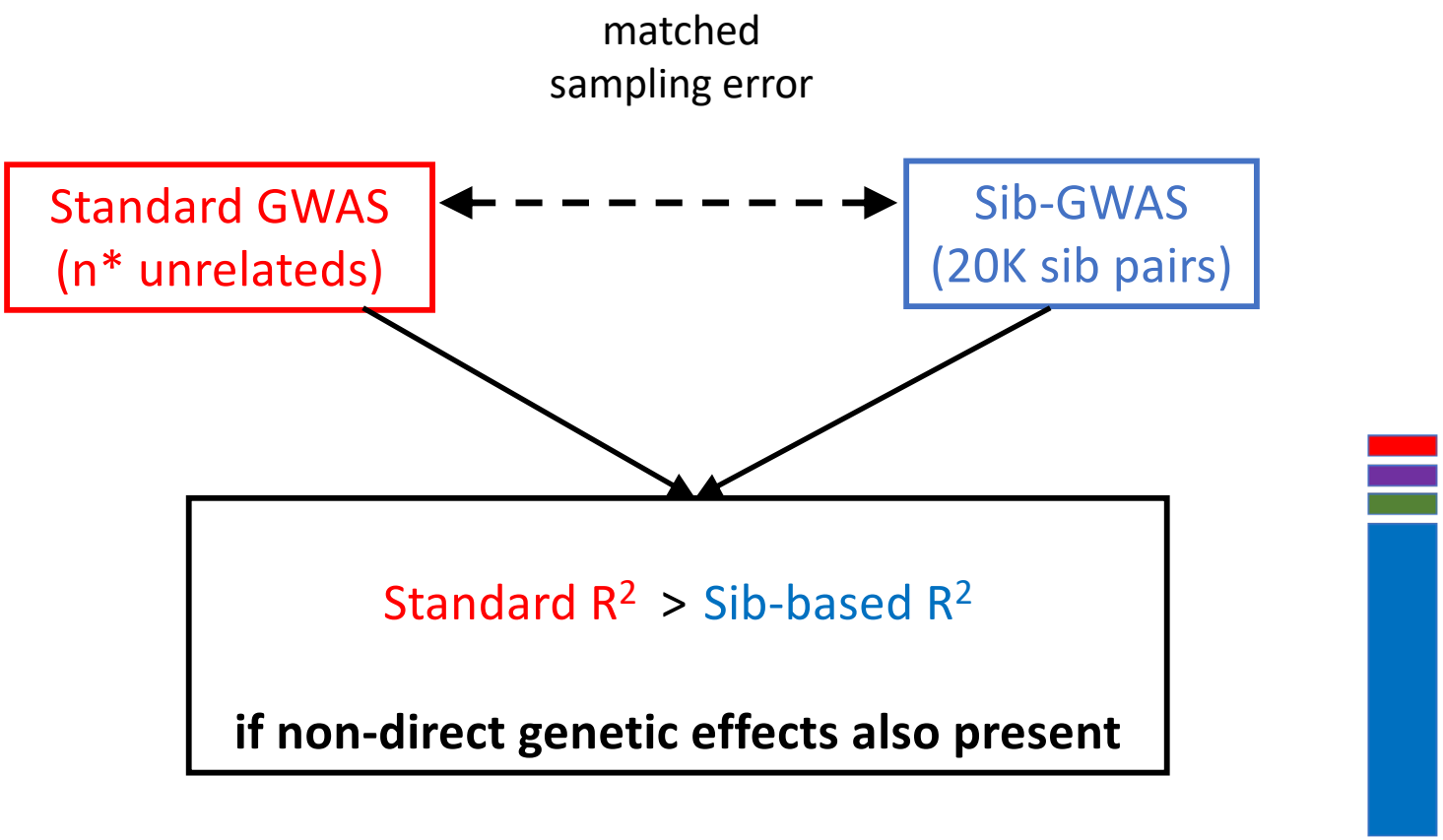
Standard GWAS: $y = \hat{\beta} x + \epsilon$

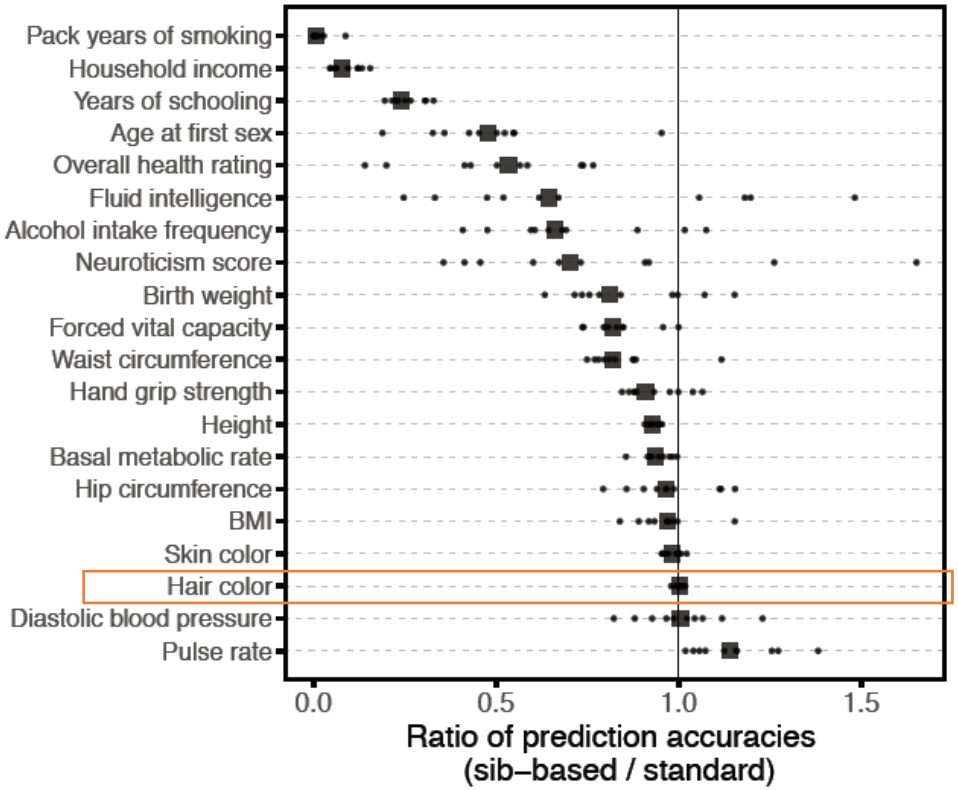
Sib-based GWAS: $\Delta y = \hat{\beta} \Delta x + \Delta \epsilon$



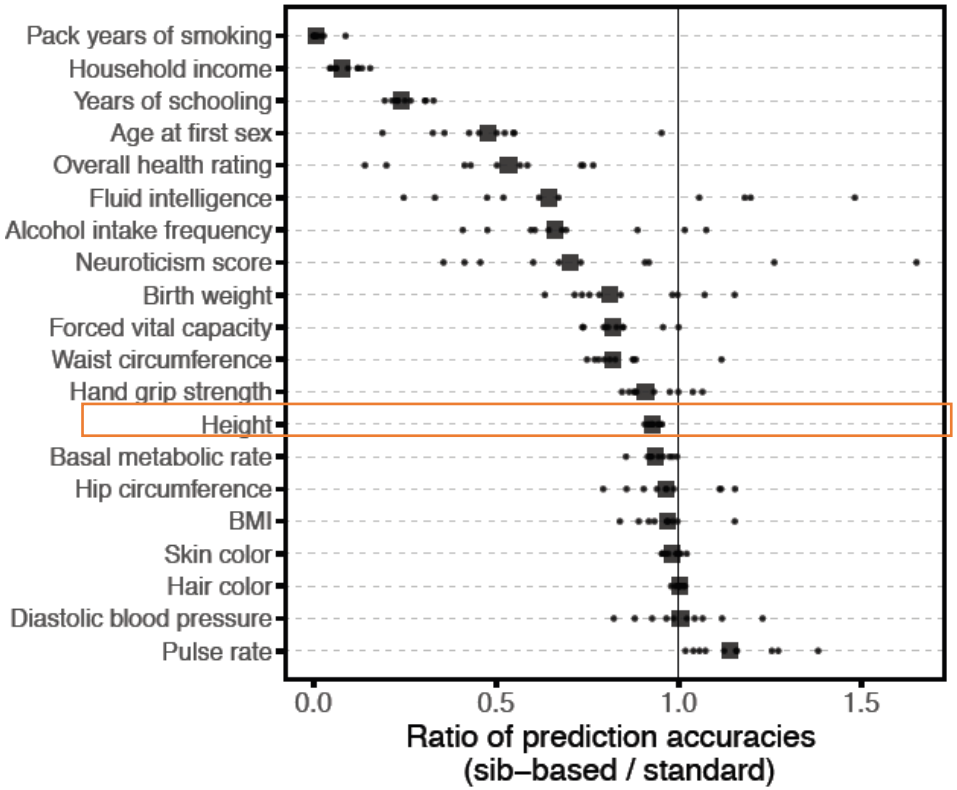

Standard GWAS
(n* unrelateds)
Sib-GWAS
(20K sib pairs)
Prediction in unrelated individuals of
same ancestries


matched
sampling error
Standard GWAS
(n* unrelateds)
Sib-GWAS
(20K sib pairs)
Standard R^2 = Sib-based R^2
Under the null

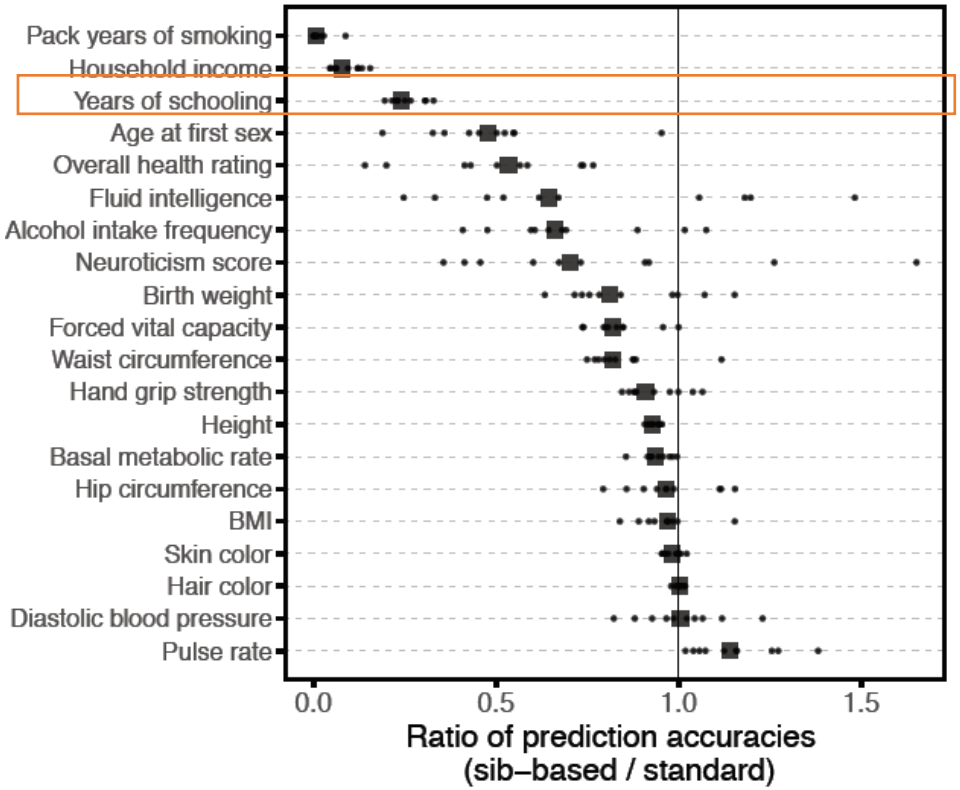

matched
sampling error
Standard GWAS
(n* unrelateds)
Sib-GWAS
(20K sib pairs)
Standard R^2 > Sib-based R^2
if non-direct genetic effects also present

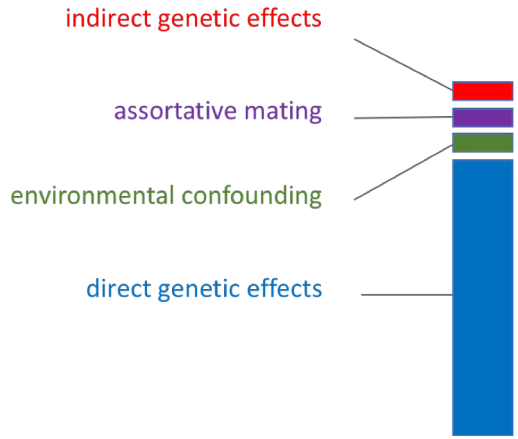


Standard GWAS
outperforms

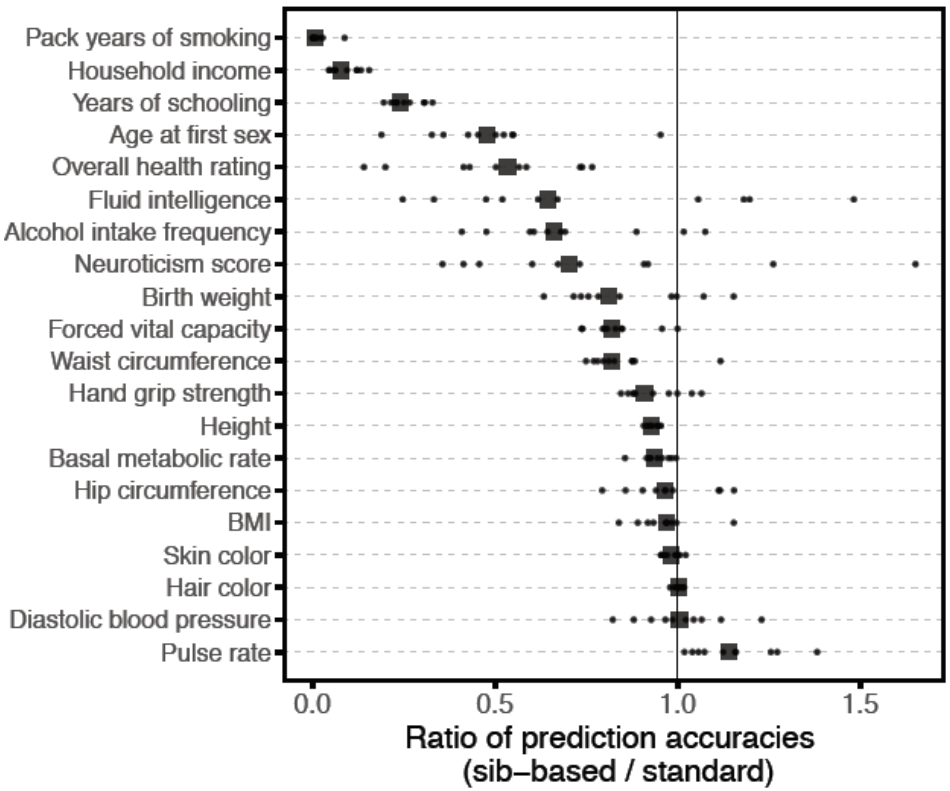


Standard GWAS
outperforms





Standard GWAS
outperforms



For many traits, PGS are not just direct genetic effects…
Do these port across cultures/environments, within an ancestry?

How do we know this is mostly about genetic ancestry?

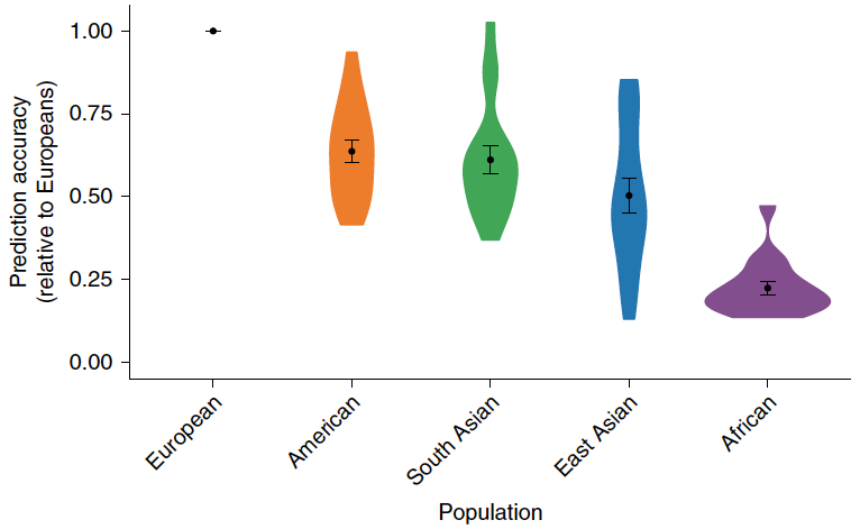


Fig. 3 | Prediction accuracy relative to European-ancestry individuals across 17 quantitative traits and 5 continental populations in the UKBB. All

Hakhamanesh Mostafavi

Arbel Harpak

Ipsita Agarwal

Dalton Conley

Jonathan Pritchard

Helpful discussions with:

Doc Edge
Guy Sella
Przeworski & Sella lab members
Graham Coop
Magnus Nordborg
Itsik Pe'er
Ziyue Gao
Augie Kong
Alex Young
Dan Belsky

NIH National Institute of General Medical Sciences

SIMONS FOUNDATION

New Results

Comment on this paper

Variable prediction accuracy of polygenic scores within an ancestry group

Hakhamanesh Mostafavi, Arbel Harpak, Dalton Conley, Jonathan K Pritchard, Molly Przeworski

doi: https://doi.org/10.1101/629949