# Hidden structure in polygenic scores and the challenge of disentangling ancestry interactions in admixed populations

Joint work with R. Mandla, Z. Shi, B. Paşaniuc and I. Mathieson

Alan J. Aw
Postdoctoral Fellow
Department of Genetics
University of Pennsylvania

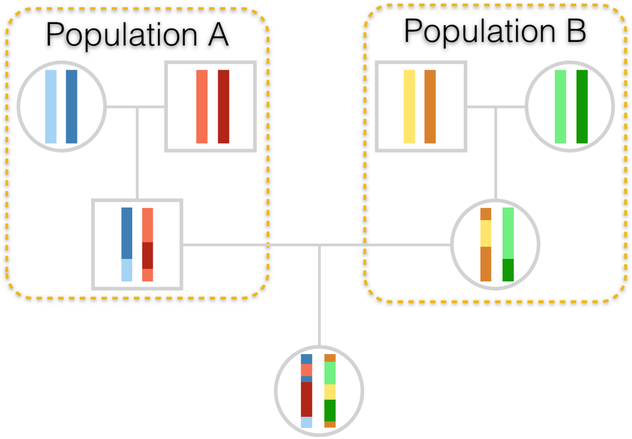



Image Source: Rodney Dyer

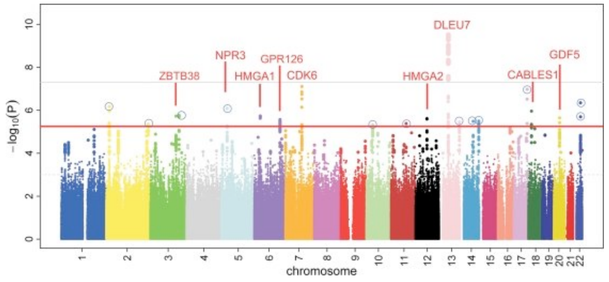
# Complex traits and poor portability

- Complex traits (e.g., height) are influenced by networks of genes that act in concert to regulate expression
- Polygenic scores trained in one population port poorly into other populations

Allelic Dosages

$$\text{PGS}_{\text{Height}} = \beta_1 x_1 + \ldots + \beta_{5000} x_{5000}$$

Variant Effect Sizes
(obtained from European GWAS,
trained on European samples)



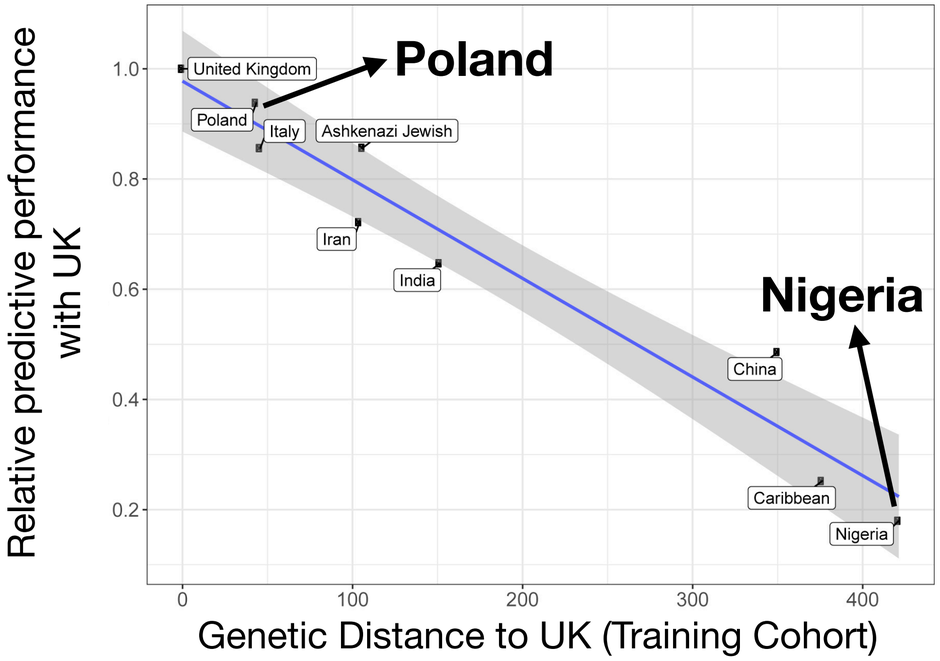




Image Sources: Estrada et al. (2009) Hum Mol Genet., Privé et al. (2022) *AJHG*

# What explains the poor portability?

1. Recent work (Hou et al., 2023 *Nat. Genet.*; Hu et al., 2025 *Nat. Genet.*) suggests high similarity in causal effects across ancestries

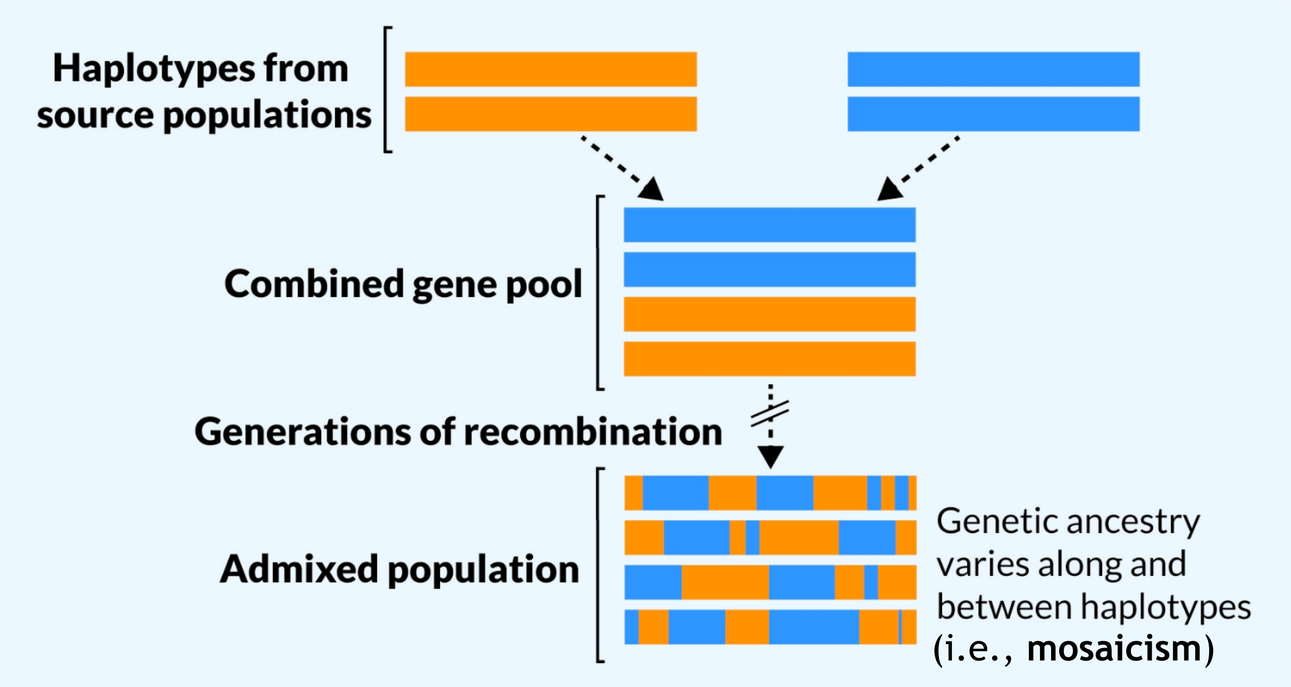
2. Differences in linkage disequilibrium (LD) patterns and allele frequencies between ancestries

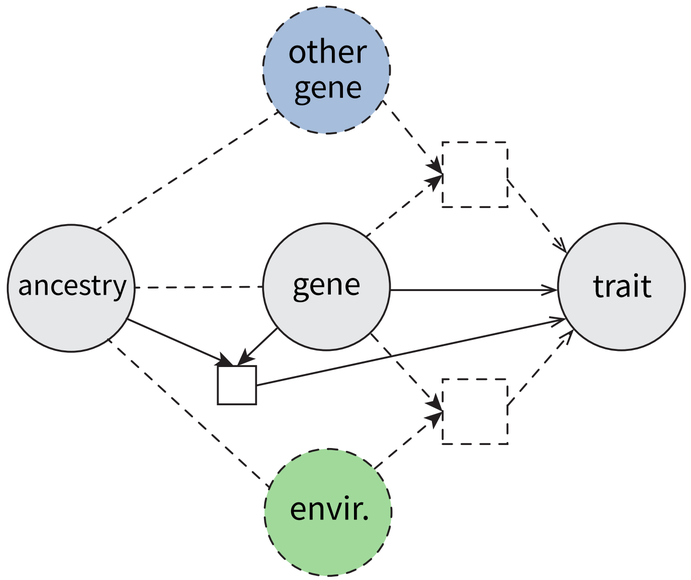
3. Interactions (Gene-by-gene [GxG] and Gene-by-environment [GxE])

   - How can causal effects be highly similar in spite of interactions?

# The role of admixed populations

- Ancestry mosaicism in admixed individuals can capture differences in allele frequencies and environmental exposures

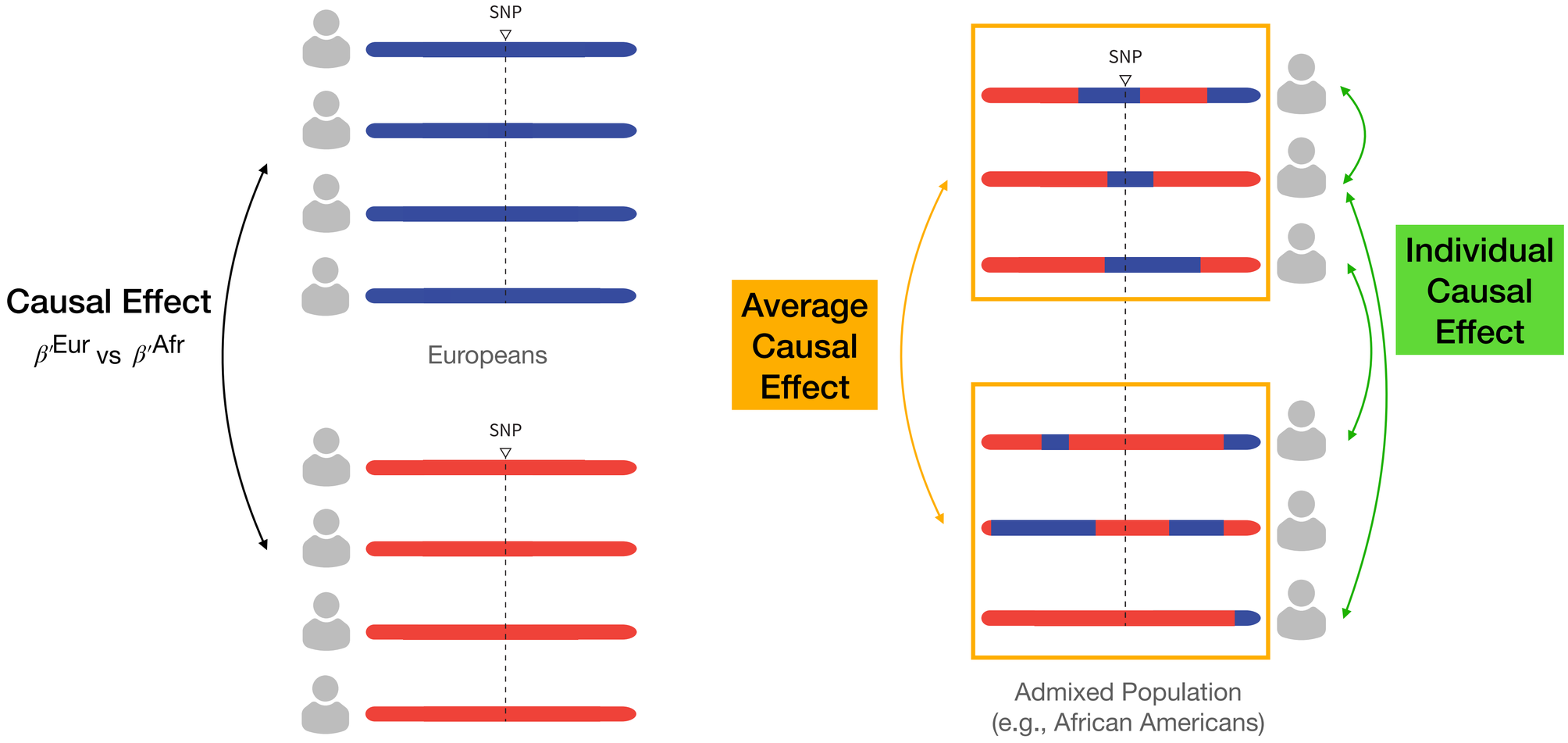



Build Statistical Models of Gene-by-Ancestry (GxA) Interactions

Image Source: Korunes and Goldberg (2021) *PLoS Genetics*

# Causal effects are similar *between what*?

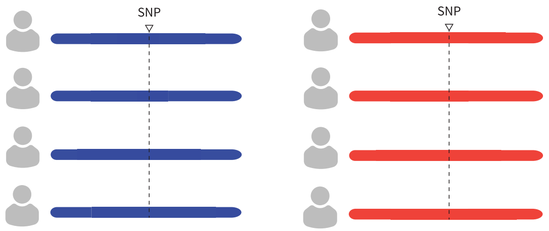



Hou et al. (2023) and Hu et al. (2025): Average causal effects are highly similar across local ancestries

# Base model of causal effects

- Ancestral non-admixed population causal effect sizes follow a bivariate normal distribution:

$$\begin{bmatrix} \beta'^{\text{Eur}} \\ \beta'^{\text{Afr}} \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma'^{2}_{\text{Eur}} & \tau' \\ \tau' & \sigma'^{2}_{\text{Afr}} \end{bmatrix} \right)$$




Variance: $$\sigma'^{2}_{\text{Eur}} = \frac{r^2}{2\sum_{j=1}^{p} f_j'^{\text{Eur}}(1 - f_j'^{\text{Eur}})} \qquad \sigma'^{2}_{\text{Afr}} = \frac{r^2}{2\sum_{j=1}^{p} f_j'^{\text{Afr}}(1 - f_j'^{\text{Afr}})}$$

Covariance: $$\tau' = \frac{\rho r^2}{2\sqrt{\sum_{j=1}^{p} f_j'^{\text{Eur}}(1 - f_j'^{\text{Eur}})}\sqrt{\sum_{j=1}^{p} f_j'^{\text{Afr}}(1 - f_j'^{\text{Afr}})}}$$

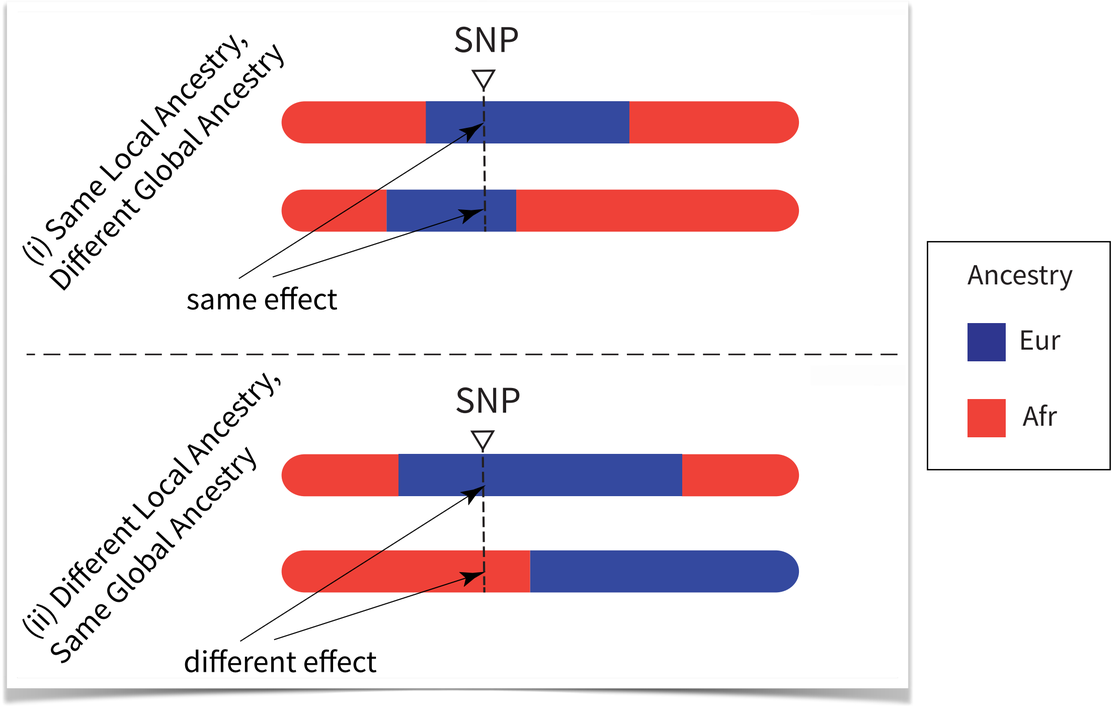
Causal effect correlation $= \tau' \Big/ \sqrt{\sigma'^{2}_{\text{Eur}}\sigma'^{2}_{\text{Afr}}} = \rho$

(Approach is similar to Wang et al., 2020 *Nat. Comm.*)

# Two models of gene-by-ancestry interaction

## Local Model

Captures interactions in *cis*

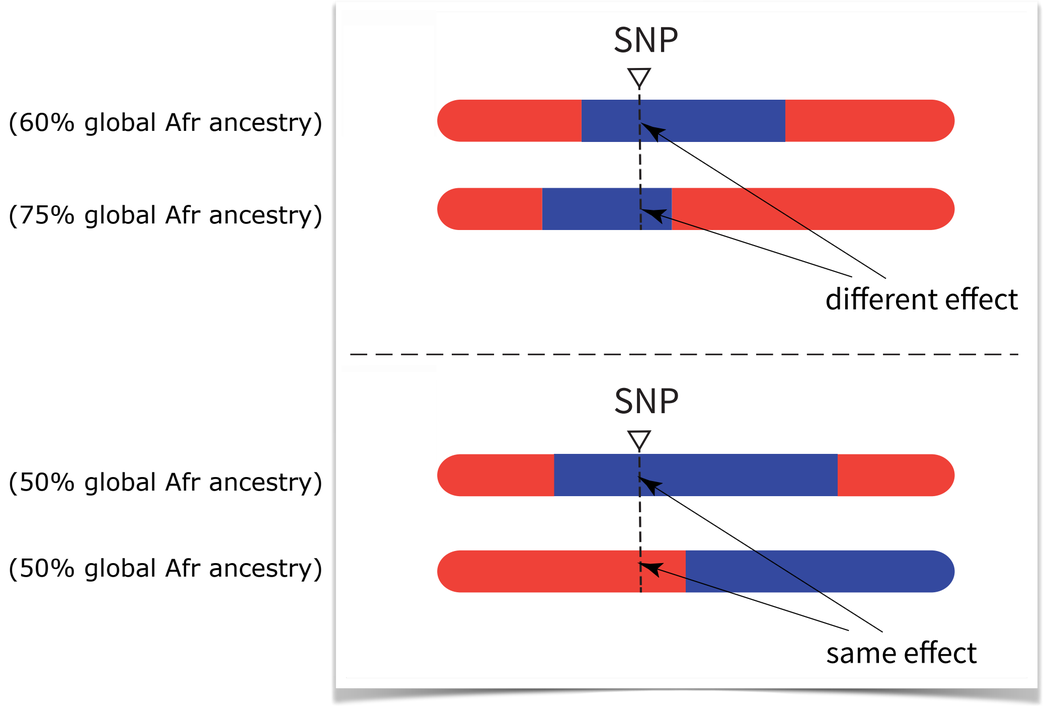



$$\beta'_{\text{SNP}} = \beta'^{\text{Afr}}\, a + \beta'^{\text{Eur}}\, (1 - a)$$

$$a = \begin{cases} 1 & \text{if ancestry is Afr} \\ 0 & \text{if ancestry is Eur} \end{cases}$$

## Global Model

Captures interactions in *trans* and GxE



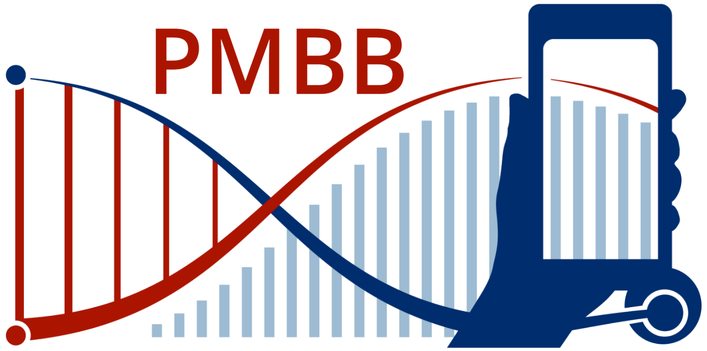

$$\beta'_{\text{SNP}} = \beta'^{\text{Afr}}\, \bar{a} + \beta'^{\text{Eur}}\, (1 - \bar{a})$$

$\bar{a}$ = genome-wide/global Afr ancestry

# Questions

1. What do the local and global models imply about individual and average causal effects?

2. Can polygenic scores differentiate the global and local models?




**Penn Medicine Biobank**

10,000 genotyped African Americans

30,000 genotyped European Americans

6 quantitative traits



Image Source: Perelman School of Medicine

# Questions

1. **What do the local and global models imply about individual and average causal effects?**

2. Can polygenic scores differentiate the global and local models?

**Penn Medicine Biobank**

10,000 genotyped African Americans

30,000 genotyped European Americans

6 quantitative traits



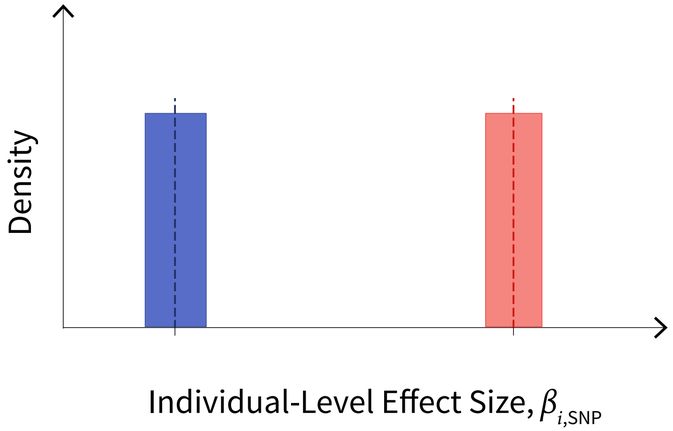
Image Source: Perelman School of Medicine

# Individual Causal Effect

## Global model implies high variability in individual effect

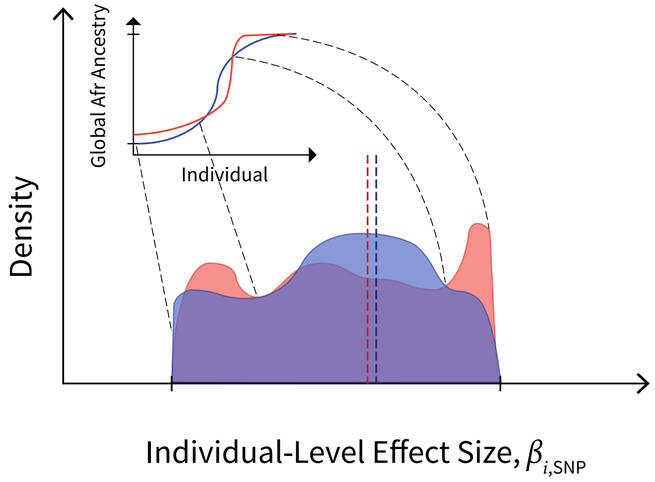
$$\beta'_{\text{SNP}} = \beta'^{\text{Afr}}\, a + \beta'^{\text{Eur}}\, (1 - a)$$

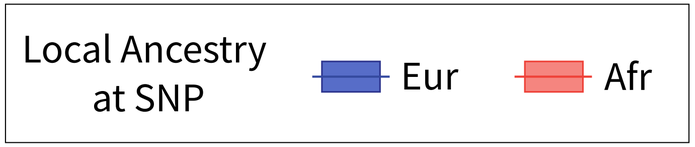
Local Model



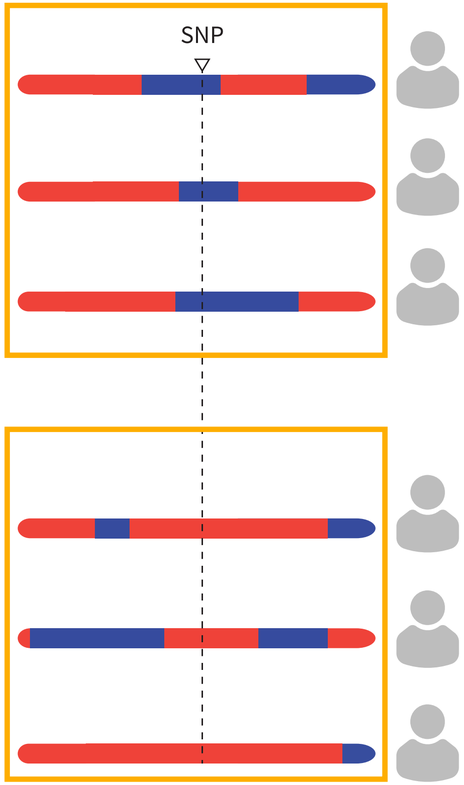

$$\beta'_{\text{SNP}} = \beta'^{\text{Afr}}\, \bar{a} + \beta'^{\text{Eur}}\, (1 - \bar{a})$$

Global Model



# Average Causal Effect — Local Model




$\overline{\beta'}|_{\text{LA=Eur}}$

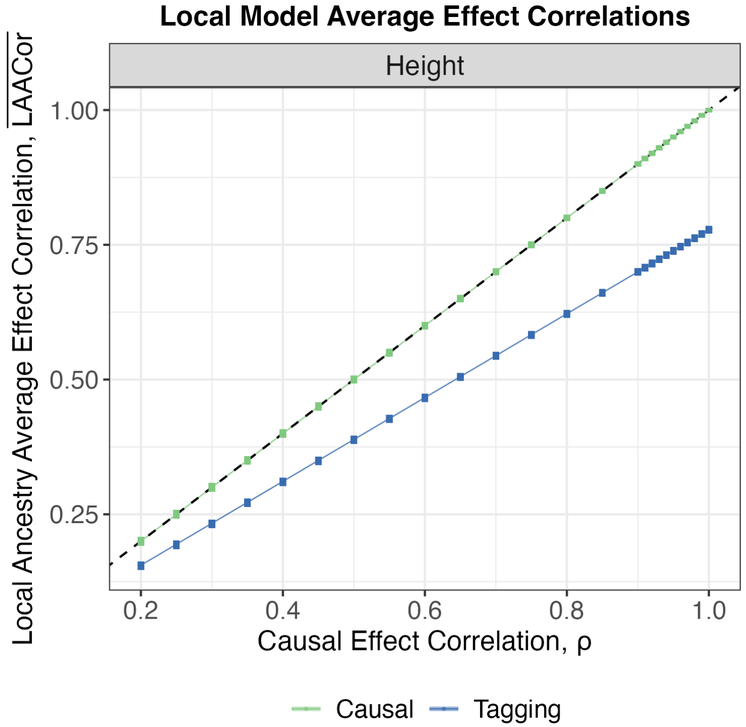
$\overline{\beta'}|_{\text{LA=Afr}}$

Under local model, distribution of average causal effect is just the distribution of causal effects in base model:

$$\begin{bmatrix} \overline{\beta'}|_{\text{LA=Eur}} \\ \overline{\beta'}|_{\text{LA=Afr}} \end{bmatrix} \overset{d}{=} \begin{bmatrix} \beta'^{\text{Eur}} \\ \beta'^{\text{Afr}} \end{bmatrix}$$

LAACor = ***L***ocal ***A***ncestry ***A***verage Causal Effect ***Cor***relation

$$\overline{\text{LAACor}'_{\text{Loc}}} \overset{def}{=} \frac{\text{cov}\left(\overline{\beta'}|_{\text{LA=Eur}}, \overline{\beta'}|_{\text{LA=Afr}}\right)}{\sqrt{\text{var}\left(\overline{\beta'}|_{\text{LA=Eur}}\right)\text{var}\left(\overline{\beta'}|_{\text{LA=Afr}}\right)}} = \rho$$



# Average Causal Effect — Global Model

## Global model produces high average causal effect similarity

**Proposition 4.2** (Joint Distribution of Local Ancestry Average Causal Effects)**.** *Under the local model, the joint distribution of average causal effects is the same as the original joint distribution of causal effects in the base model, Eq. (2). Let* $\beta'_{\cdot j}|_{LA=Afr}$ *and* $\beta'_{\cdot j}|_{LA=Eur}$ *denote the African and European local ancestry average causal effects under the global model. The joint distribution of these quantities is*

$$\begin{bmatrix}\beta'_{\cdot j}|_{LA=Afr}\\ \beta'_{\cdot j}|_{LA=Eur}\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}u'_j & w'_j\\ w'_j & v'_j\end{bmatrix}\right),$$

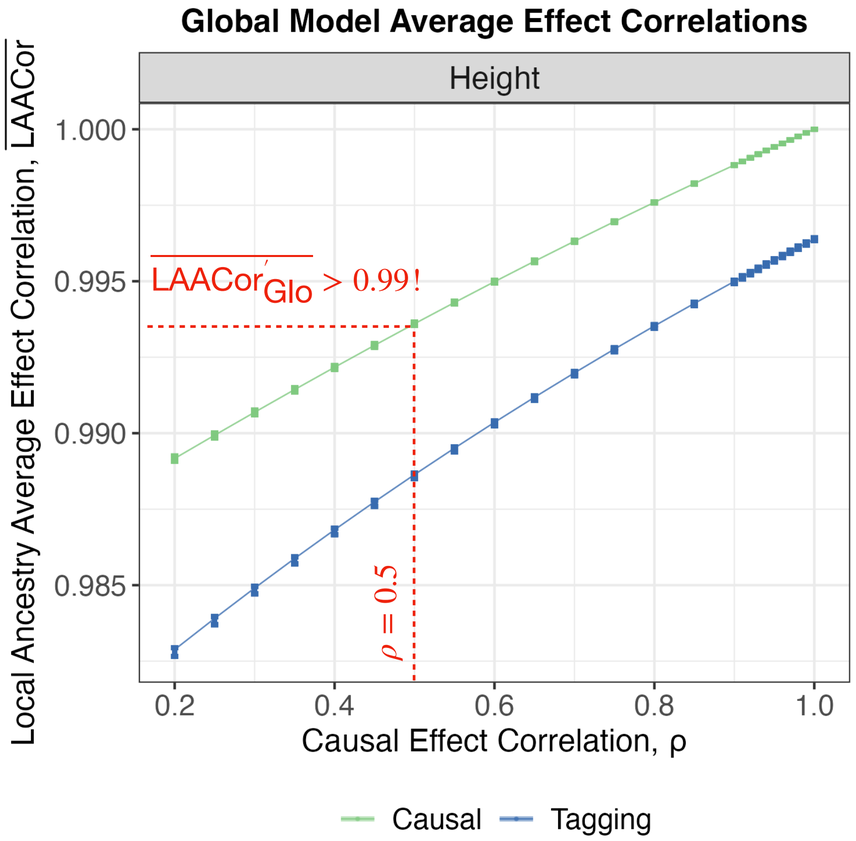
*where*

$$\begin{aligned} u'_j &= \sigma'^2_{Eur}\omega'^2_{1,j} + 2\tau'\omega'_{1,j}\omega'_{2,j} + \sigma'^2_{Afr}\omega'^2_{2,j}\\ v'_j &= \sigma'^2_{Eur}\omega'^2_{3,j} + 2\tau'\omega'_{3,j}\omega'_{4,j} + \sigma'^2_{Afr}\omega'^2_{4,j}\\ w'_j &= \sigma'^2_{Eur}\omega'_{1,j}\omega'_{3,j} + \tau'(\omega'_{2,j}\omega'_{3,j} + \omega'_{1,j}\omega'_{4,j}) + \sigma'^2_{Afr}\omega'_{2,j}\omega'_{4,j} \end{aligned}$$

*are terms in the covariance matrix, with quantities* $\sigma'^2_{Afr}, \sigma'^2_{Eur}$ *and* $\tau'$ *defined in Eqs. (3)-(5), and quantities* $\omega'_{1,j}, \omega'_{2,j}, \omega'_{3,j}, \omega'_{4,j}$ *defined in Supplementary Material Subsection S8 (Box C) depending only on the haplotype and local ancestry matrices.*

LAACor = ***L***ocal ***A***ncestry ***A***verage Causal Effect ***Cor***relation

$$\implies \overline{\text{LAACor}'_{\text{Glo}}} \approx \frac{\sum_{j=1}^{p} w'_j}{\sqrt{\sum_{j=1}^{p} u'_j}\sqrt{\sum_{j=1}^{p} v'_j}}$$



# Summary of Q1

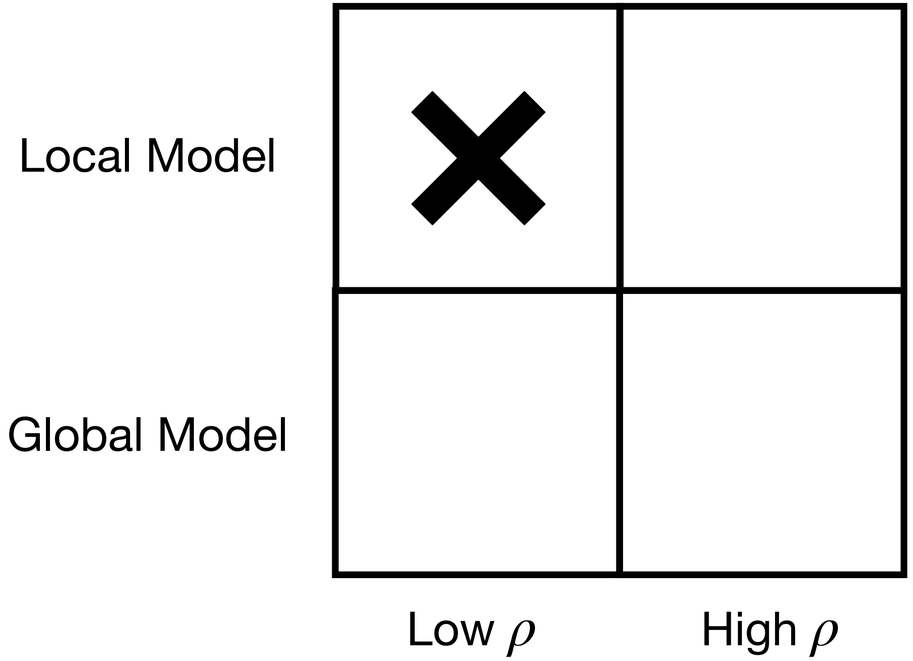
- High (average) causal effect similarity does not rule out variability in individual causal effect

| Model | Individual Causal Effect Variability | Average Causal Effect Similarity |
|---|---|---|
| Local Model | ⬇️ | Same as causal effect correlation $\rho$ |
| Global Model | ⬆️ | Can be very high, despite small $\rho$ |

**Example: Height**

Hou et al. (2023): LAACor is about 0.94

Hu et al. (2025): LAACor lies in [0.9, 1], 95% CI contains 1



# Questions

1. What do the local and global models imply about individual and average causal effects?

2. **Can polygenic scores differentiate the global and local models?**

**Penn Medicine Biobank**

10,000 genotyped African Americans

30,000 genotyped European Americans

6 quantitative traits



Image Source: Perelman School of Medicine

# Polygenic scores (computed on tagging variants)

Causal Effect

$$\beta^{\text{Eur}} = \beta'^{\text{Eur}} \cdot \text{LD}^{\text{Eur}} \cdot \sqrt{\frac{f'^{\text{Eur}}\left(1 - f'^{\text{Eur}}\right)}{f^{\text{Eur}}\left(1 - f^{\text{Eur}}\right)}}$$

Causal and Tagging Variant LD

Causal Variant Allele Frequency

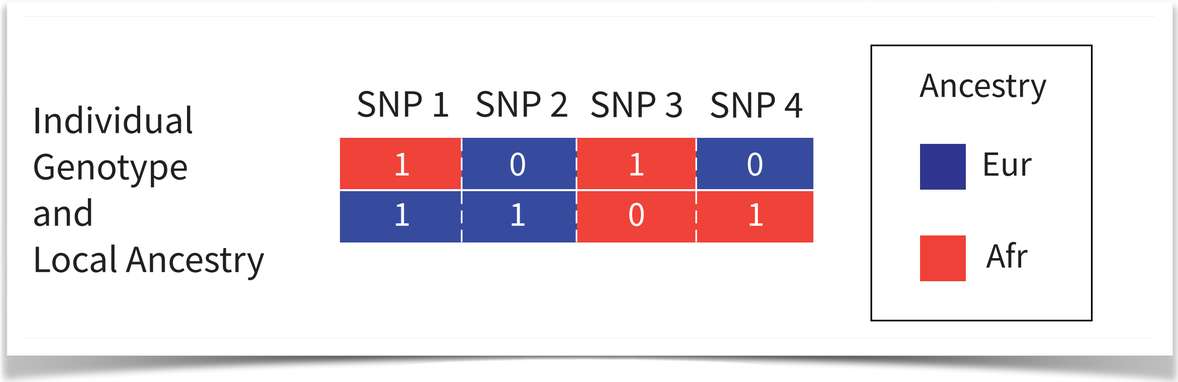
Tagging Variant Allele Frequency

$$\beta^{\text{Afr}} = \beta'^{\text{Afr}} \cdot \text{LD}^{\text{Afr}} \cdot \sqrt{\frac{f'^{\text{Afr}}\left(1 - f'^{\text{Afr}}\right)}{f^{\text{Afr}}\left(1 - f^{\text{Afr}}\right)}}$$

See: Zaidi (2020) [blog post]; Vukcevic et al. (2011) *AJHG*

# Polygenic scores (computed on tagging variants)

- **Standard, or Total, polygenic score:** assign European effect sizes to all alleles
- **Partial polygenic score:** restrict to genomic chunks of European ancestry only




$$\text{TotPGS} = \beta_1 \times \left( \begin{array}{c} 1 - f_1^{\text{Afr}} \\ + \\ 1 - f_1^{\text{Eur}} \end{array} \right) + \beta_2 \times \left( \begin{array}{c} -f_2^{\text{Eur}} \\ + \\ 1 - f_2^{\text{Eur}} \end{array} \right) + \beta_3 \times \left( \begin{array}{c} 1 - f_3^{\text{Afr}} \\ + \\ -f_3^{\text{Afr}} \end{array} \right) + \beta_4 \times \left( \begin{array}{c} -f_4^{\text{Eur}} \\ + \\ 1 - f_4^{\text{Afr}} \end{array} \right)$$

$$\text{ParPGS} = \beta_1 \times \left( 1 - f_1^{\text{Eur}} \right) + \beta_2 \times \left( \begin{array}{c} -f_2^{\text{Eur}} \\ + \\ 1 - f_2^{\text{Eur}} \end{array} \right) + \beta_3 \times \left( \quad \right) + \beta_4 \times \left( -f_4^{\text{Eur}} \right)$$

(More on Partial PGS: Sun et al., 2024 *Nat. Comm.*; Marnetto et al., 2020 *Nat. Comm.*; Bitarello and Mathieson, 2020 *G3*)

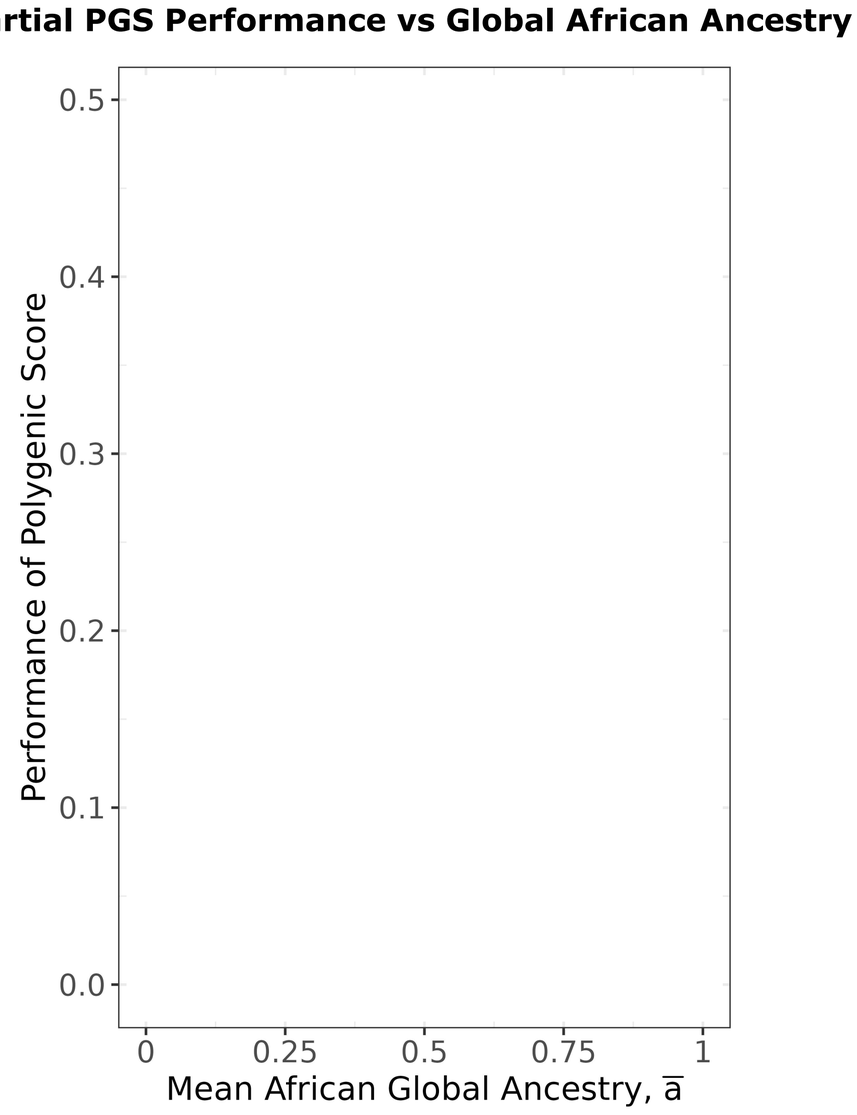
# Causal Variants Known: Partial PGS differentiates the two models (but Total PGS does not)

- Partial PGS performance declines **cubically** in global ancestry under the global model, but declines **linearly** under the local model

Correlation in causal effects between ancestries

**Global Model:** $\mathbb{E}[\text{cor}^2(ParPGS, y)] \approx r^2(1-\overline{a})(1-\overline{a}+\rho\overline{a})^2$

(Mean African) Global ancestry



# Causal Variants Known: Partial PGS differentiates the two models (but Total PGS does not)

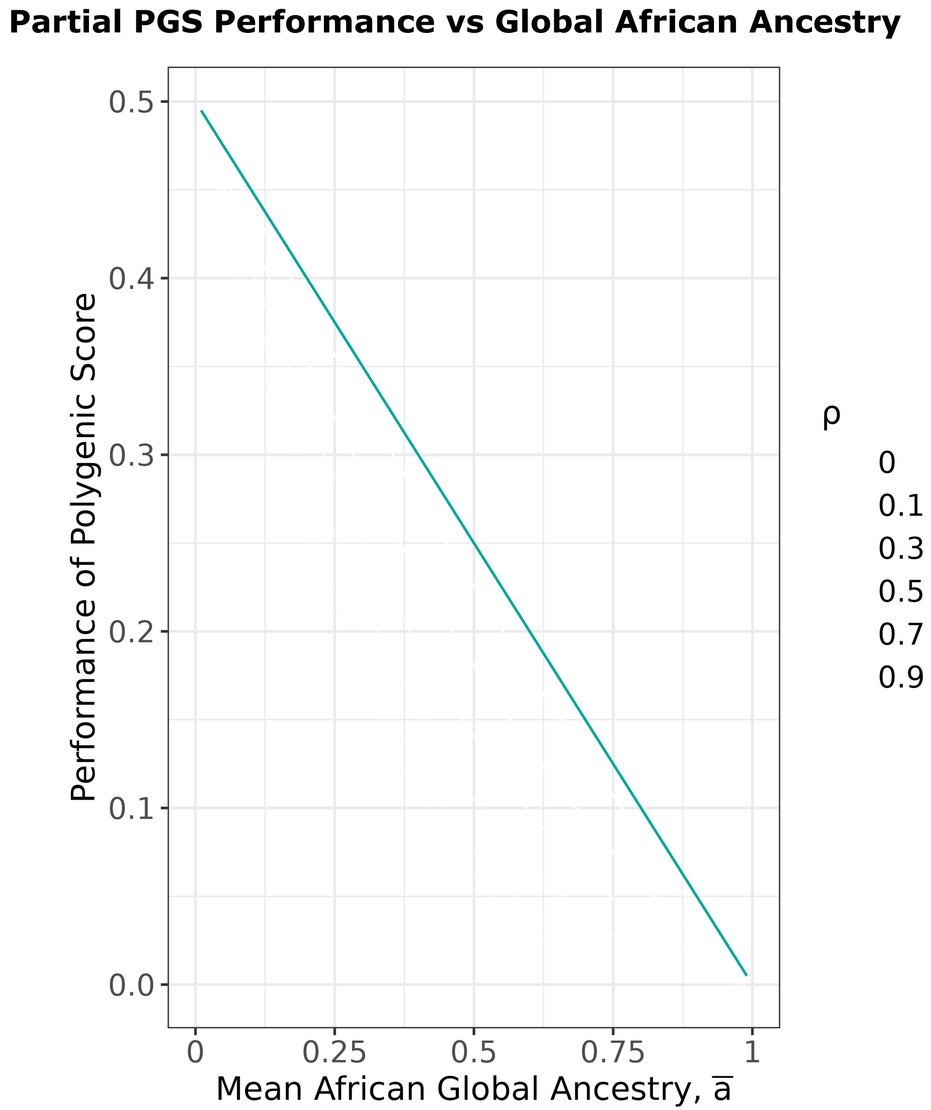
- Partial PGS performance declines **cubically** in global ancestry under the global model, but declines **linearly** under the local model

Correlation in tagging effects between ancestries

**Global Model:** $\mathbb{E}[\text{cor}^2(ParPGS, y)] \approx r^2(1-\bar{a})(1-\bar{a}+\rho\bar{a})^2$

**Local Model:** $\mathbb{E}[\text{cor}^2(ParPGS, y)] \approx r^2(1-\bar{a})$

(Mean African) Global ancestry



# Causal Variants Known: Partial PGS differentiates the two models (but Total PGS does not)

- Partial PGS performance declines **cubically** in global ancestry under the global model, but declines **linearly** under the local model

Correlation in causal effects between ancestries

**Global Model:** $\mathbb{E}[\text{cor}^2(ParPGS, y)] \approx r^2(1-\bar{a})(1-\bar{a}+\rho\bar{a})^2$

**Local Model:** $\mathbb{E}[\text{cor}^2(ParPGS, y)] \approx r^2(1-\bar{a})$

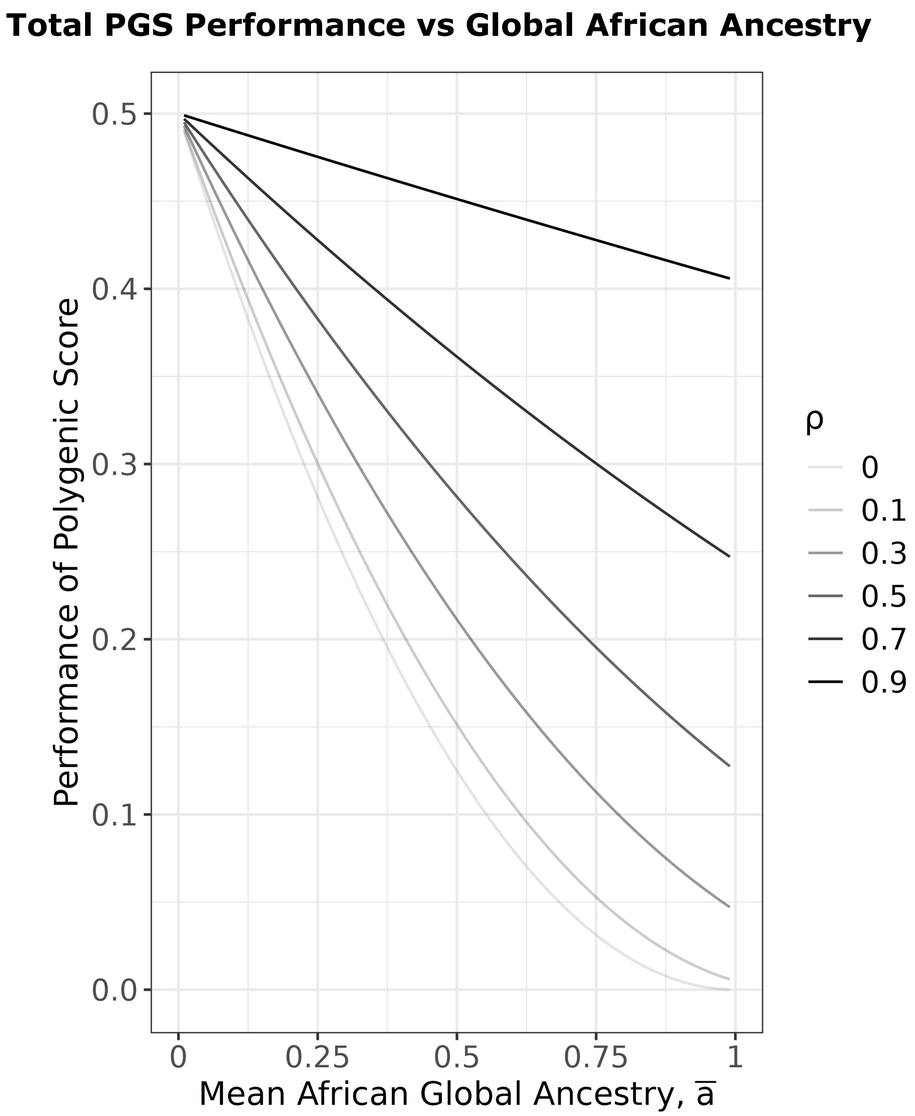
(Mean African) Global ancestry

- Total PGS performance declines **quadratically** in global ancestry under either model

$$\mathbb{E}[\text{cor}^2(TotPGS, y)] \approx r^2(1-\bar{a}+\rho\bar{a})^2$$

Correlation in causal effects between ancestries

(Mean African) Global ancestry



# Causal Variants Known: Partial PGS differentiates the two models (but Total PGS does not)

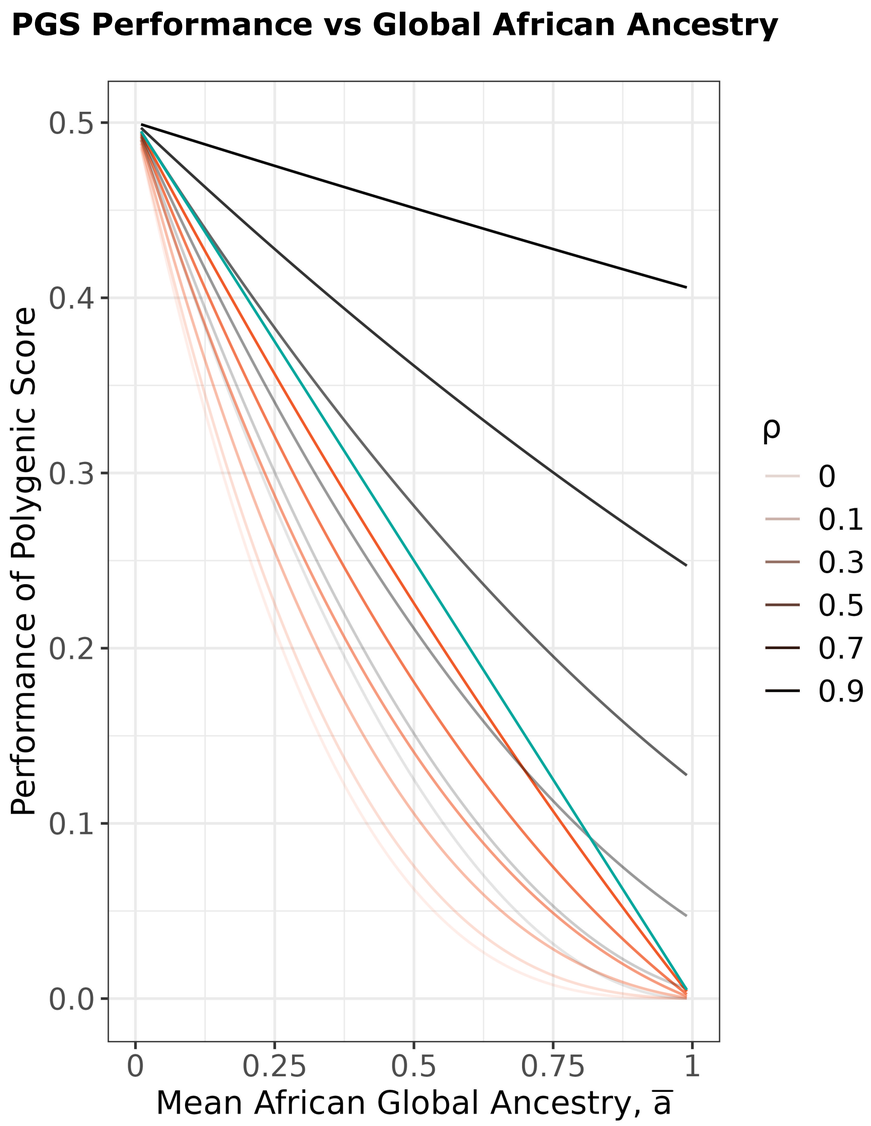
- Partial PGS performance declines **cubically** in global ancestry under the global model, but declines **linearly** under the local model

**Global Model:** $\mathbb{E}[\text{cor}^2(ParPGS, y)] \approx r^2(1-\bar{a})(1-\bar{a}+\rho\bar{a})^2$

**Local Model:** $\mathbb{E}[\text{cor}^2(ParPGS, y)] \approx r^2(1-\bar{a})$

- Total PGS performance declines **quadratically** in global ancestry under either model

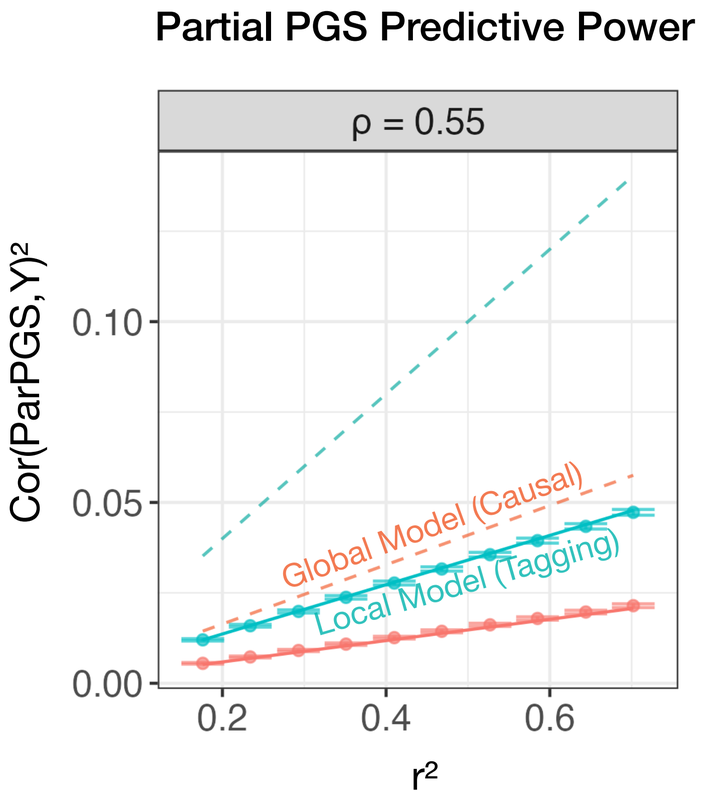
$$\mathbb{E}[\text{cor}^2(TotPGS, y)] \approx r^2(1-\bar{a}+\rho\bar{a})^2$$



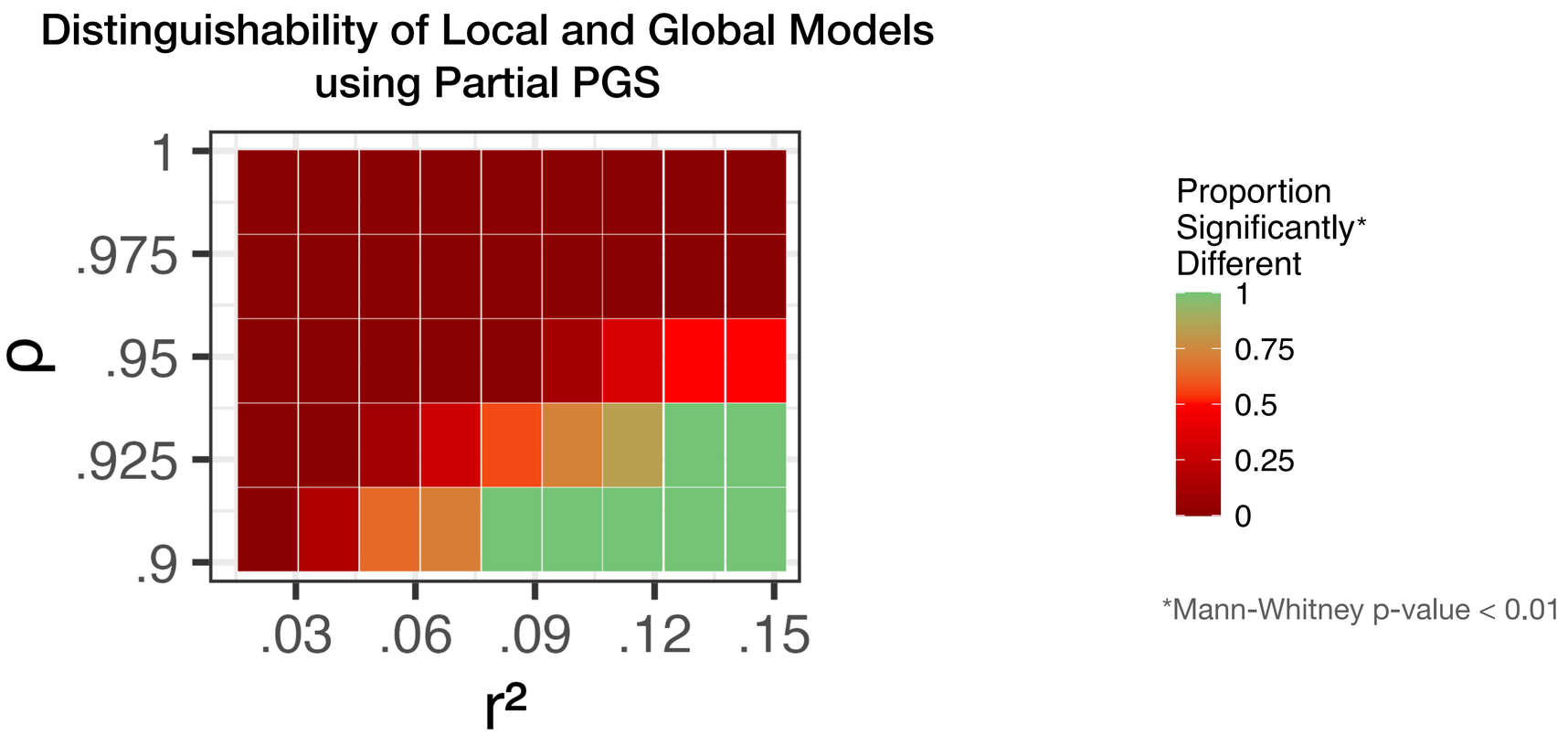
# Causal Variants Unknown: Heterogeneity in LD and allele frequencies hinders differentiation of models

Both $\mathbb{E}[\text{cor}(TotPGS, y)^2]$ and $\mathbb{E}[\text{cor}(ParPGS, y)^2]$ depend on causal-tagging LD and causal allele frequencies

LD and causal AF heterogeneities may produce differences in the two models that resemble analytical differences



# High causal effect correlation also hinders distinguishability of local and global models in general



# Summary of Q2

## Causal Variants Known (Ideal)

- Can differentiate local and global models using ParPGS

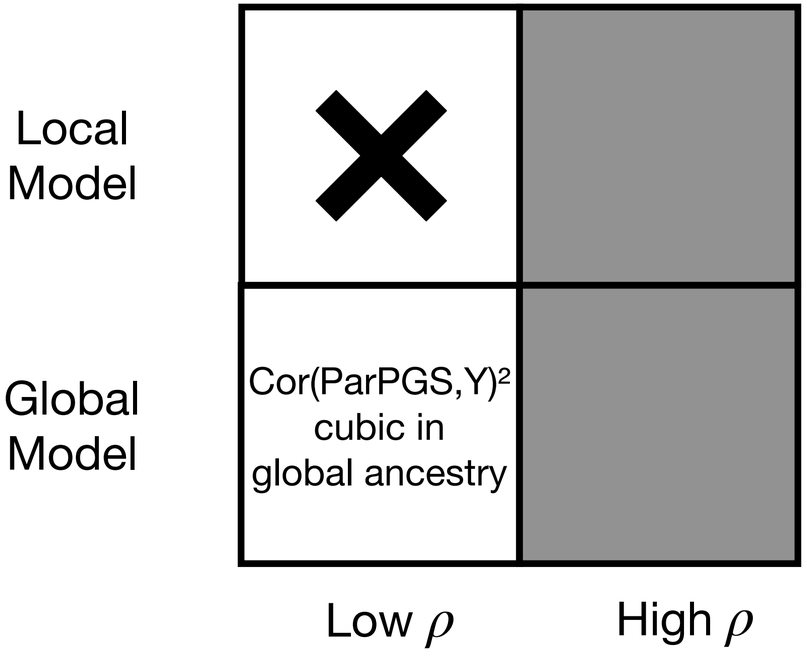
## Causal Variants Unknown (Realistic)

- Unknown differences in LD patterns and allele frequencies hinder differentiation

## High $\rho$?

- Difficult to differentiate local and global models

**Example: Height**




Assuming all polygenic score variants are causal:

50% contribution of global model

# Conclusion

- Models of GxA interaction are consistent with:
  - poor cross-ancestry portability
  - high (average) causal effect similarity across ancestries
- Fine-mapping causal variants helps differentiate the two models in future work