Gene-ancestry interactions mediate individual heterogeneity in causal effects on complex traits in admixed populations

Alan J. Aw[1], Penn Medicine Biobank and Iain Mathieson[1]
[1]Department of Genetics, Perelman School of Medicine, University of Pennsylvania
alan.aw(at)pennmedicine.upenn.edu

1. Introduction

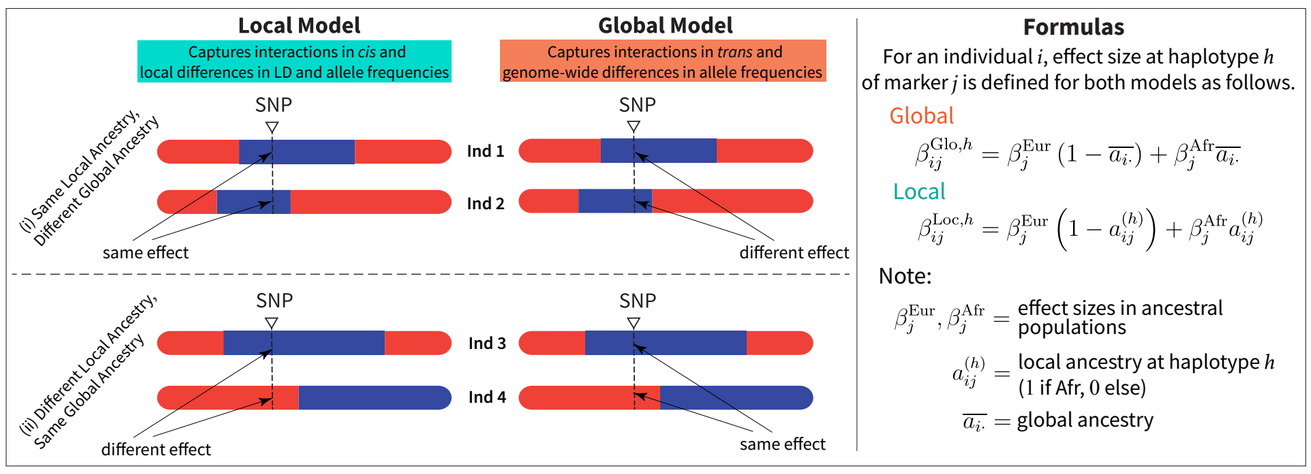
- Admixed populations are under-represented in genomic studies
- Polygenic scores (PGSs) trained on single ancestry cohorts underperform on admixed individuals
  1. Strategies to improve performance include modeling ancestry-specific effects and ancestry-specific PGSs
  2. But recent work also suggests causal effects are highly correlated across ancestries
  3. Gene-by-Gene and Gene-by-Environment interactions can also induce marginal effect heterogeneity
- Ancestry mosaicism in admixed individuals accumulates genome-wide differences in allele frequencies

Key Questions

1. How much does *trans* epistasis contribute to the genetic architecture of complex traits?
2. What does the contribution suggest about polygenic risk prediction in admixed individuals?

2. Theory

- Two models of Gene-by-Ancestry interaction:



Formulas

For an individual i, effect size at haplotype h of marker j is defined for both models as follows.

Global

$$\beta_{ij}^{\text{Glo},h} = \beta_j^{\text{Eur}}\left(1-\overline{a_{i\cdot}}\right) + \beta_j^{\text{Afr}}\overline{a_{i\cdot}}$$

Local

$$\beta_{ij}^{\text{Loc},h} = \beta_j^{\text{Eur}}\left(1-a_{ij}^{(h)}\right) + \beta_j^{\text{Afr}}a_{ij}^{(h)}$$

Note:

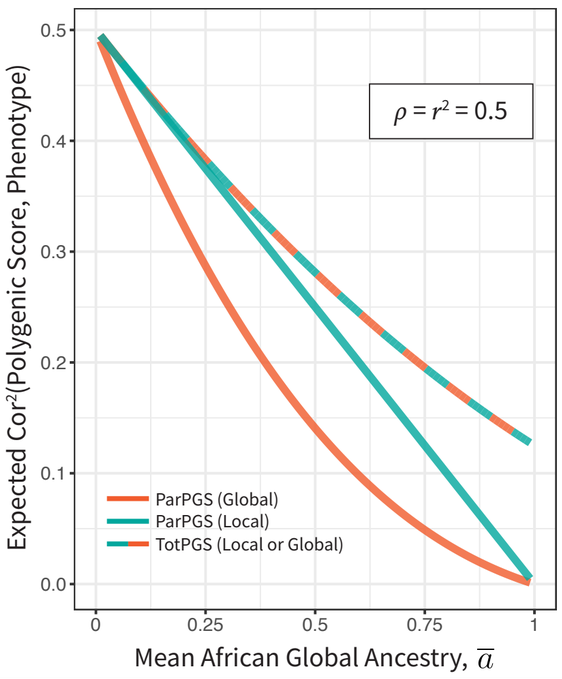
$\beta_j^{\text{Eur}}, \beta_j^{\text{Afr}}$ = effect sizes in ancestral populations

$a_{ij}^{(h)}$ = local ancestry at haplotype h (1 if Afr, 0 else)

$\overline{a_{i\cdot}}$ = global ancestry

- We computed two types of polygenic scores to investigate the fit of models to complex traits
  - **Standard, or Total, polygenic scores (TotPGS)**, which assign training cohort (European ancestry) effect sizes to all alleles: $\text{TotPGS}(\boldsymbol{x}_i, \boldsymbol{a}_i) = \sum_{j=1}^{p} \beta_j^{\text{Eur}}\left(\hat{x}_{ij}^{(1)} + \hat{x}_{ij}^{(2)}\right)$, where $\boldsymbol{x}_i$ is the genotype
  - **Partial polygenic scores (ParPGS)**, which restrict scores to genomic chunks of training cohort ancestry only: $\text{ParPGS}(\boldsymbol{x}_i, \boldsymbol{a}_i) = \sum_{j=1}^{p} \beta_j^{\text{Eur}}\left[\left(1-a_{ij}^{(1)}\right)\hat{x}_{ij}^{(1)} + \left(1-a_{ij}^{(2)}\right)\hat{x}_{ij}^{(2)}\right]$
- ParPGS differentiates local and global models, but TotPGS does not

PGS Performance vs Global African Ancestry



Mathematical Results

Underlying Assumptions (Base Model)

$$\begin{bmatrix}\beta_j^{\text{Eur}}\\ \beta_j^{\text{Afr}}\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma_{\text{Eur}}^2 & \tau\\ \tau & \sigma_{\text{Afr}}^2\end{bmatrix}\right) \qquad (\star)$$

$$\sigma_{\text{Eur}}^2 = \frac{r^2}{2\sum_{j=1}^{p} \hat{f}_j^{\text{Eur}}(1-\hat{f}_j^{\text{Eur}})}$$

$$\sigma_{\text{Afr}}^2 = \frac{r^2}{2\sum_{j=1}^{p} \hat{f}_j^{\text{Afr}}(1-\hat{f}_j^{\text{Afr}})}$$

$$\tau = \frac{r^2\rho}{2\sqrt{\sum_{j=1}^{p} \hat{f}_j^{\text{Eur}}(1-\hat{f}_j^{\text{Eur}})}\sqrt{\sum_{j=1}^{p} \hat{f}_j^{\text{Afr}}(1-\hat{f}_j^{\text{Afr}})}}$$

Under ($\star$), the following equations describe the performance of PGSs under the Local and Global Models

ParPGS

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

Local $\quad \mathbb{E}[\text{cor}^2(\text{ParPGS}, y)] \approx r^2(1-\overline{a})$

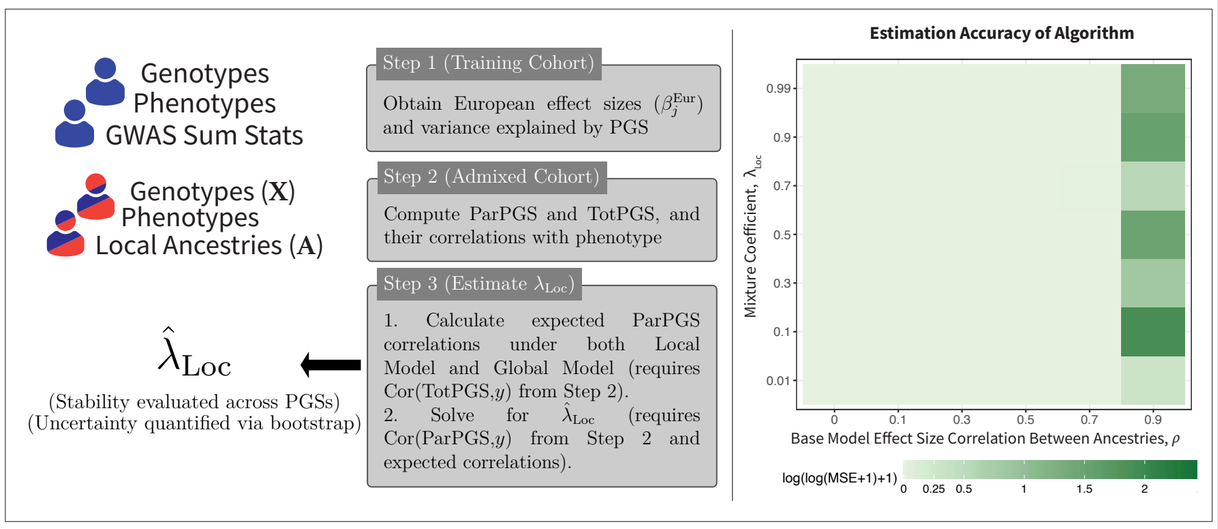
TotPGS

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

3. Methods

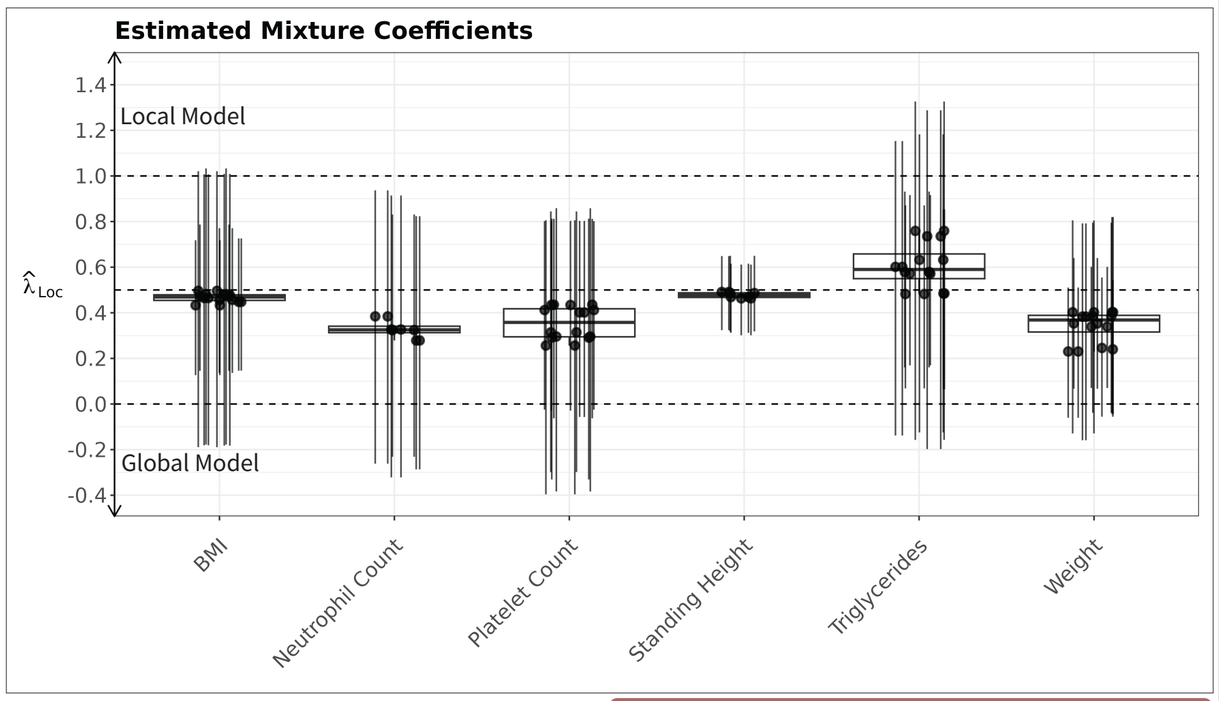
- Both the global model and the local model contribute to complex trait architecture, so we introduce **mixture models**
  - Effect size is convex combination of effect sizes under the global and local models: $\beta_{ij}^{(h)} = \lambda_{\text{Loc}}\beta_{ij}^{\text{Loc},h} + (1-\lambda_{\text{Loc}})\beta_{ij}^{\text{Glo},h}$

- We develop a "moment matching"-like approach to estimate λ_{Loc} from polygenic scores, and establish its accuracy via simulations

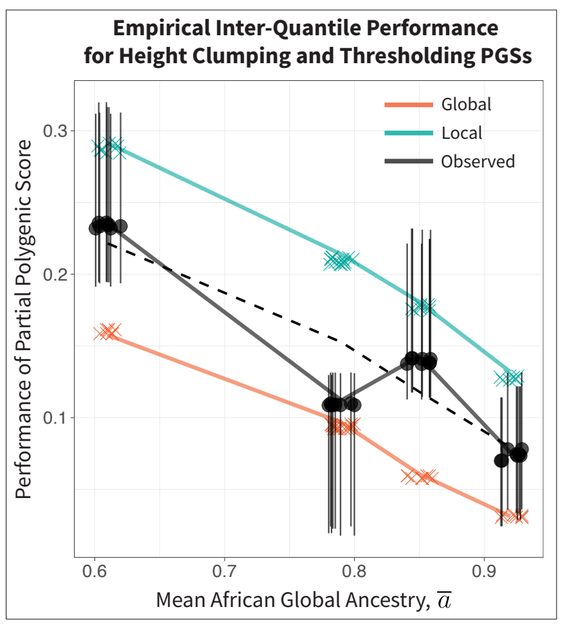


4. Application to Penn Medicine Biobank (PMBB)

- Using UK Biobank summary statistics, perform Clumping and Thresholding (C&T) on $n = 29{,}410$ unrelated PMBB participants of largely European ancestry to construct PGSs
- Local ancestry of $n = 9{,}324$ PMBB participants of mixed African and European ancestry (**ADM**) inferred using RFMix
- Run estimation approach on ADM to obtain $\hat{\lambda}_{\text{Loc}}$



Empirical Inter-Quantile Performance for Height Clumping and Thresholding PGSs



Significance of Results

We demonstrate the existence of substantial gene-by-ancestry interactions in complex traits. Such interactions introduce individual-level effect heterogeneity, but are consistent with highly similar average causal effects (which was recently reported in Hou et al., 2023 *Nat. Genet.* and Hu et al., 2023 *bioRxiv*). Our findings are inconsistent with local ancestry entirely explaining differences in predictive power of polygenic scores. Therefore, polygenic scores for admixed individuals should include both local and global ancestry.

Acknowledgements: We thank Regeneron Genetics Center for help with data generation, and the National Institutes of Health for funding this work.