Genetic association and heritability estimation in structured populations

Alejandro Ochoa

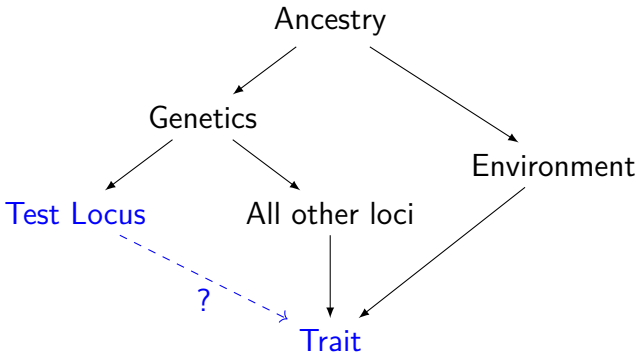
StatGen, Biostatistics & Bioinformatics — Duke University

2023-11-18 — CBB retreat

Part 1: Cryptic relatedness matters a lot in genetic association studies ... or: how I learned to love Linear Mixed-effects Models

Association studies are hard

- Millions of tests
- Polygenicity (many causal variants)
- Incorrect assumptions: independence / additivity
- **Confounders**



PCA vs LMM in association

Principal Components Analysis (PCA) and Linear Mixed-effects Model (LMM):

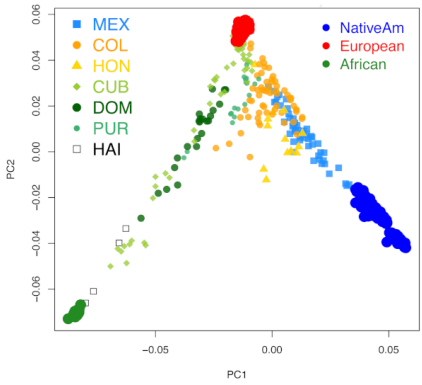
$$\text{PCA}: \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{U}_d\gamma_d + \epsilon,$$

$$\text{LMM}: \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{s} + \epsilon.$$

$\mathbf{U}_d$ are top d eigenvectors of kinship matrix Φ.

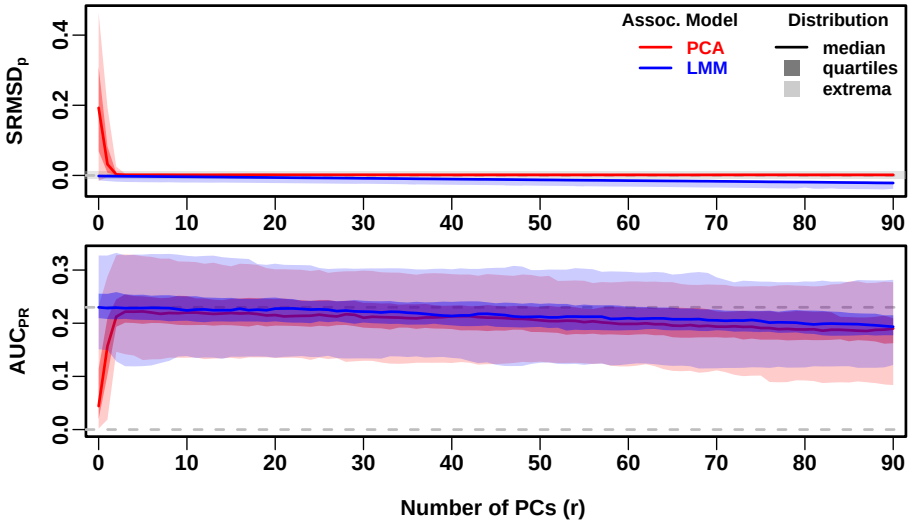
$$\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, \sigma_G^2\Phi\right).$$

- PCA is faster but low-dimensional
- LMM is slower but can model families



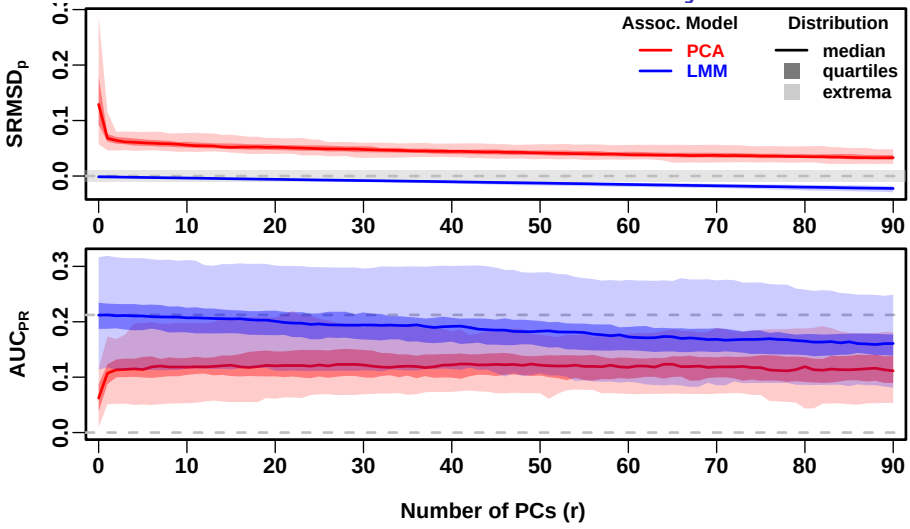
Moreno-Estrada *et al.* (2013)

PCA < LMM in association: simulated admixture



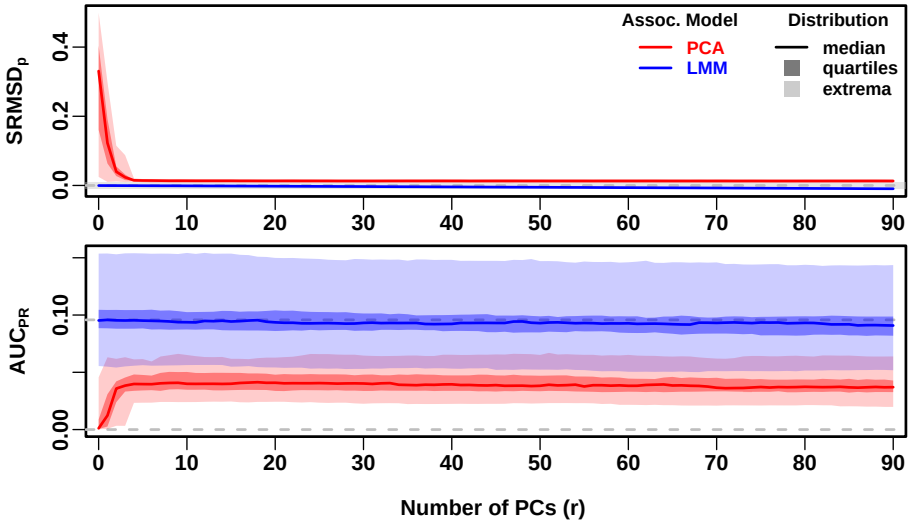
Yao and Ochoa (2023) doi:10.7554/eLife.79238

PCA < LMM in association: simulated family structure



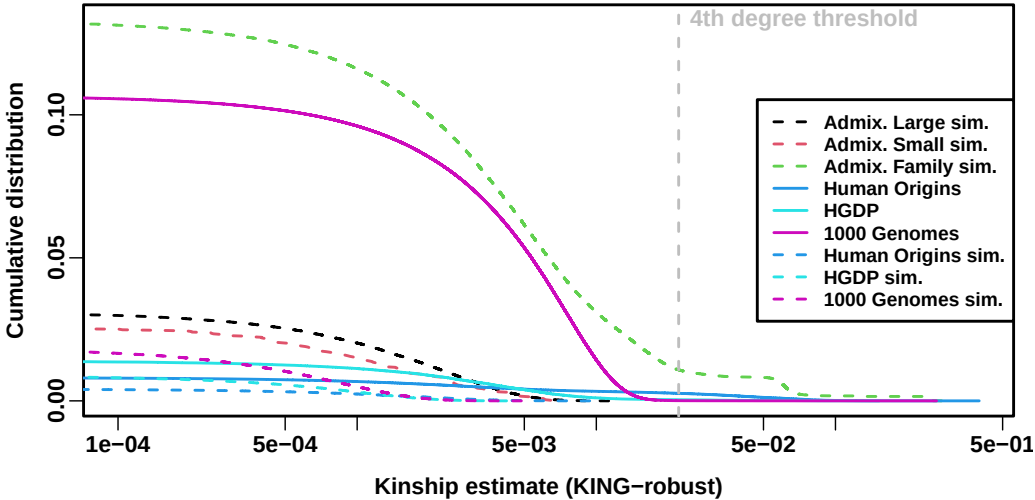
Yao and Ochoa (2023) doi:10.7554/eLife.79238

PCA < LMM in association: 1000 Genomes



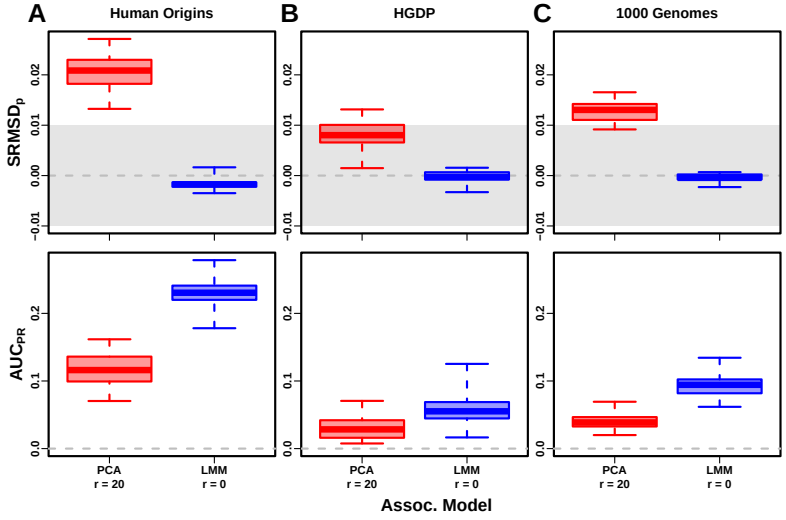
Yao and Ochoa (2023) doi:10.7554/eLife.79238

Numerous distant relatives in real datasets



Yao and Ochoa (2023) doi:10.7554/eLife.79238

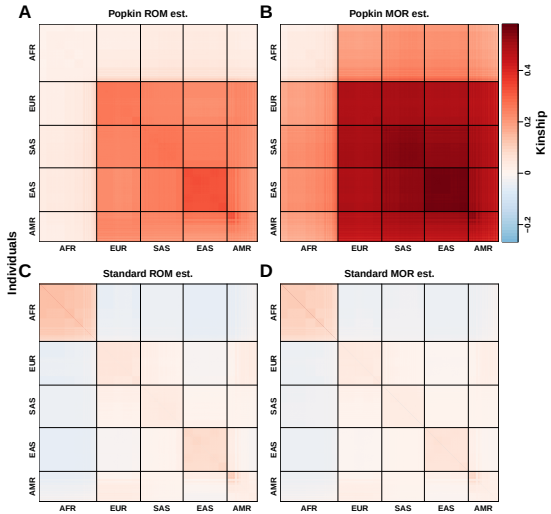
Numerous distant relatives in real datasets explain PCA < LMM



Yao and Ochoa (2023) doi:10.7554/eLife.79238

Part 2: Kinship bias carries over to heritability, but not association!

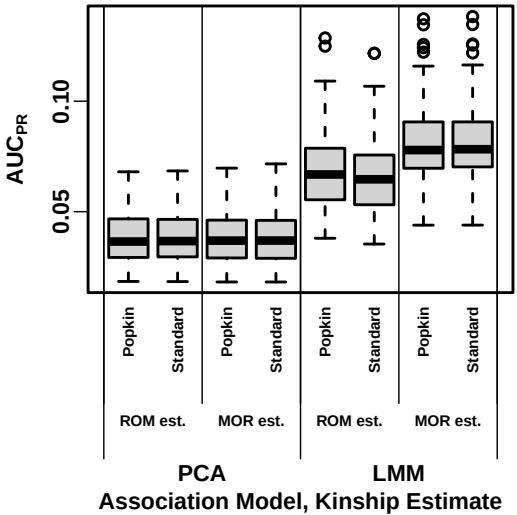
Standard kinship estimator is severely biased



- ROM: Ratio of Means
  - Upweighs common variants
  - Behaves well statistically
- MOR: Mean of Ratios
  - Upweighs rare variants
  - Introduces additional bias

Hou and Ochoa (2023) doi:10.1093/genetics/iyad030

Kinship bias does not affect genetic associations



Hou and Ochoa (2023) doi:10.1093/genetics/iyad030

Kinship bias does not affect genetic associations

Linear algebra proof!

Transforming true to biased kinship matrices:

Φ : Unbiased kinship matrix,

Φ' : Biased kinship matrix,

$$\Phi' = \frac{1}{1-\bar{\varphi}}\mathbf{C}\Phi\mathbf{C},$$

$\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$: Centering matrix.

Kinship bias does not affect genetic associations

Linear algebra proof!

Transforming true to biased kinship matrices:

Φ : Unbiased kinship matrix,

Φ' : Biased kinship matrix,

$$\Phi' = \frac{1}{1-\bar{\varphi}}\mathbf{C}\Phi\mathbf{C},$$

$\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$: Centering matrix.

Association test is a regression with correlated residuals:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{s} + \epsilon,$$
$$\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, 2\sigma^2_G\Phi\right),$$
$$\epsilon \sim \text{Normal}\left(\mathbf{0}, \sigma^2_E\mathbf{I}\right).$$

Kinship bias does not affect genetic associations

Linear algebra proof!

Transforming true to biased kinship matrices:

Φ : Unbiased kinship matrix,

Φ' : Biased kinship matrix,

$$\Phi' = \frac{1}{1-\bar{\varphi}}\mathbf{C}\Phi\mathbf{C},$$

$\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$: Centering matrix.

Association test is a regression with correlated residuals:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{s} + \epsilon,$$
$$\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, 2\sigma^2_G\Phi\right),$$
$$\epsilon \sim \text{Normal}\left(\mathbf{0}, \sigma^2_E\mathbf{I}\right).$$

Kinship bias compensated by intercept!

$$\mathbf{s}' = \mathbf{C}\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, 2\sigma^{2\prime}_G\Phi'\right),$$
$$\sigma^{2\prime}_G = (1-\bar{\varphi})\sigma^2_G,$$
$$\mathbf{s}' = \mathbf{s} - \mathbf{1}\bar{s},$$
$$\alpha' = \alpha + \bar{s}$$

Hou and Ochoa (2023) doi:10.1093/genetics/iyad030

Kinship bias affects heritability estimation

Model:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{s} + \epsilon,$$
$$\mathbf{s} + \epsilon \sim \text{Normal}\left(\mathbf{0}, 2\sigma_G^2\Phi + \sigma_E^2\mathbf{I}\right).$$

Heritability definition:

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}.$$

Variance is estimated with bias:

$$\sigma_G^{2\prime} = (1 - \bar{\varphi})\sigma_G^2.$$

Kinship bias affects heritability estimation

Model:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{s} + \epsilon,$$
$$\mathbf{s} + \epsilon \sim \text{Normal}\left(\mathbf{0}, 2\sigma_G^2\Phi + \sigma_E^2\mathbf{I}\right).$$

Heritability definition:

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}.$$

Variance is estimated with bias:

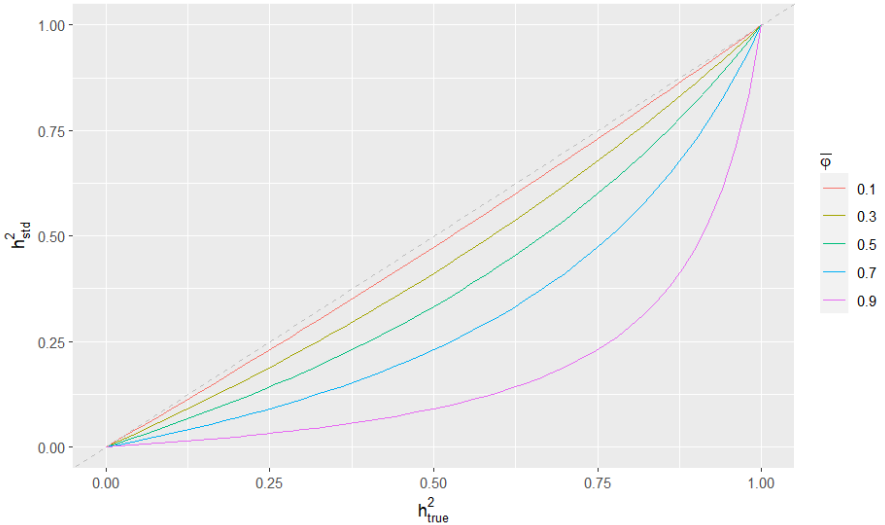
$$\sigma_G^{2\prime} = (1 - \bar{\varphi})\sigma_G^2.$$

Heritability is estimated with bias that depends on mean kinship $\bar{\varphi}$ and the true heritability h^2:

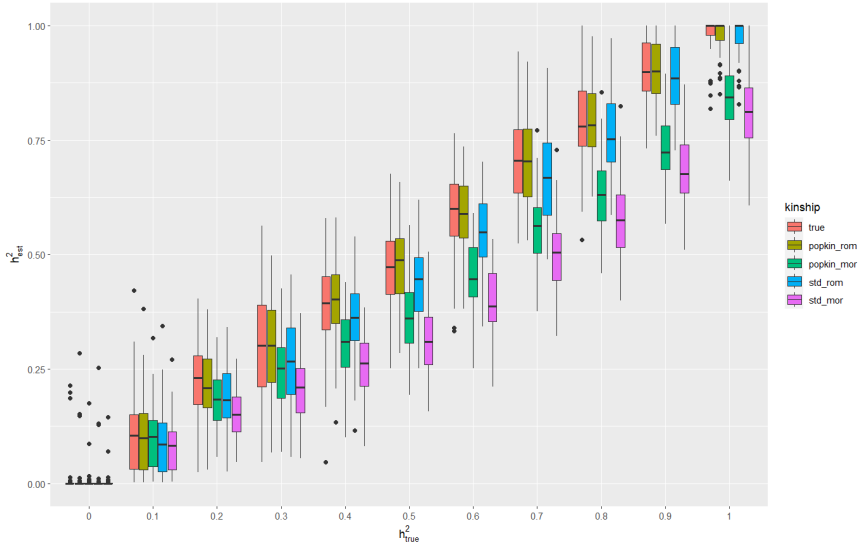
$$\begin{aligned} h^{2\prime} &= \frac{\sigma_G^{2\prime}}{\sigma_G^{2\prime} + \sigma_E^{2\prime}} \\ &= h^2 \frac{1 - \bar{\varphi}}{1 - \bar{\varphi}h^2}. \end{aligned}$$

Kinship bias affects heritability estimation

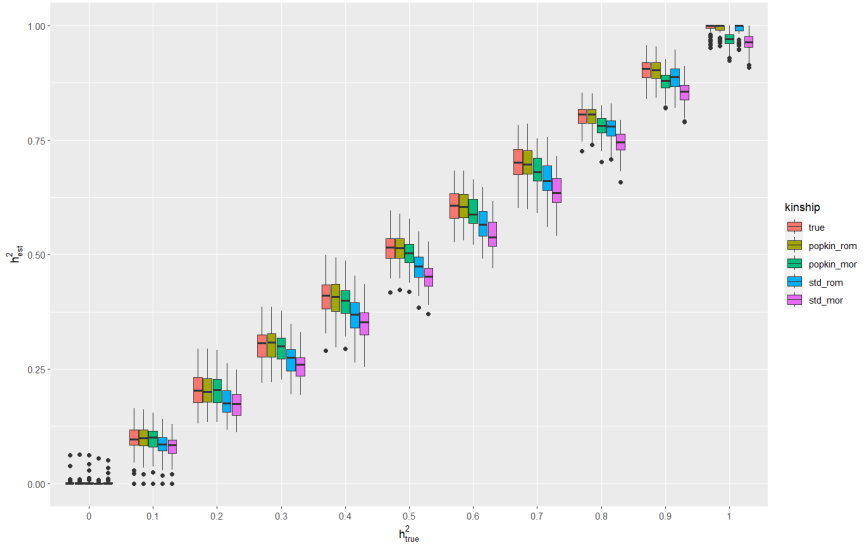
$$h^{2\prime} = h^2 \frac{1-\bar{\varphi}}{1-\bar{\varphi}h^2}$$



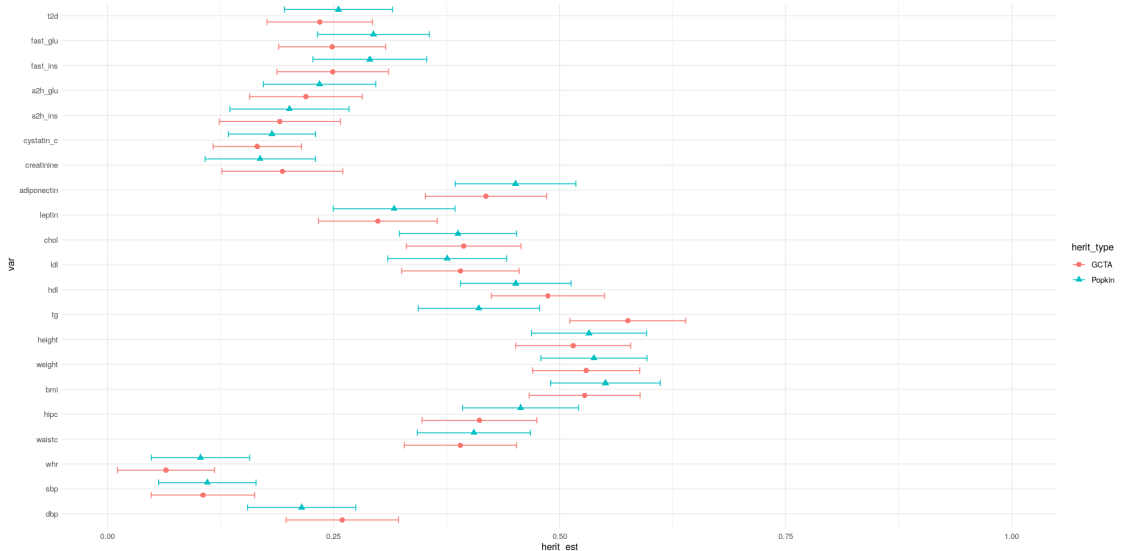
Simulated admixture (high mean kinship)



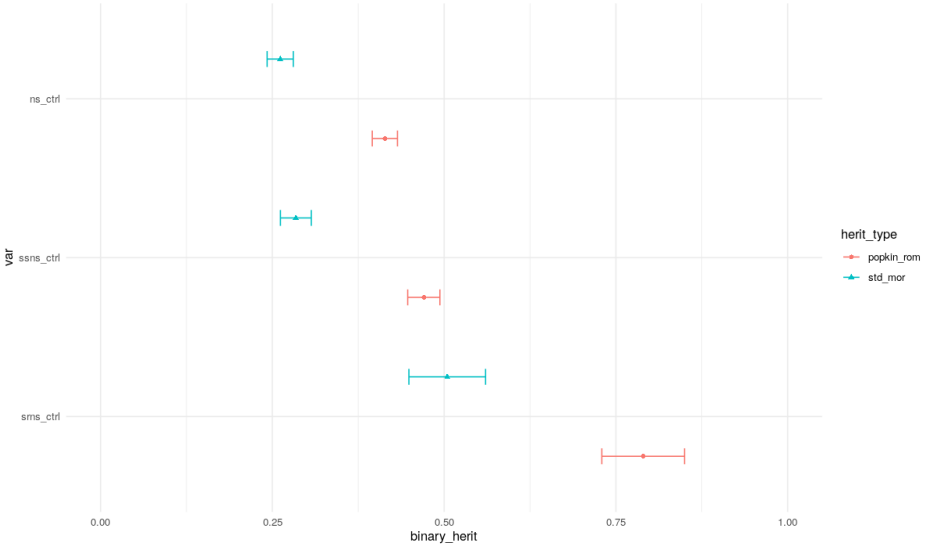
Simulated family structure (reduced mean kinship)



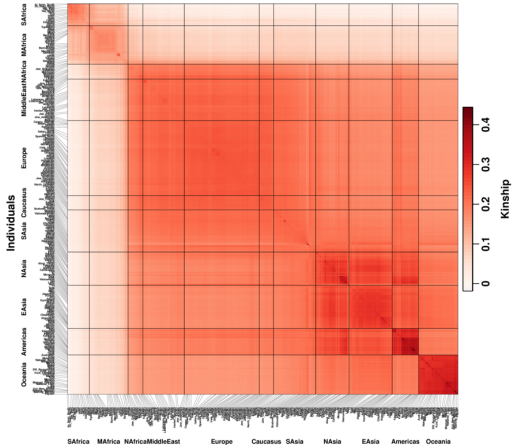
San Antonio Family Study: Type 2 Diabetes (low mean kinship)



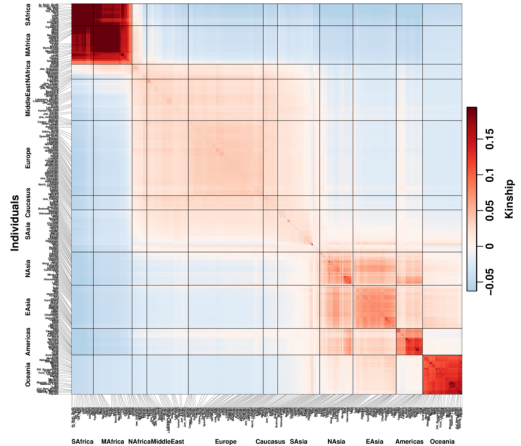
Nephrotic Syndrome: multiethnic cohort (high mean kinship)



Unbiased kinship estimates: new models, opportunities



New "popkin"
kinship estimator



Biased "standard"
kinship estimator

Acknowledgments

Ochoa Lab
Tiffany Tu
RP Pornmongkolsuk
Zhuoran Hou
Yiqi Yao
Amika Sood
Danielle Mensah

Princeton University
John D. Storey

Duke University
Rasheed Gbadegesin
Kouros Owzar
Beth Hauser
Yi-Ju Li
Andrew Allen
Amy Goldberg

Funding
NIH
Whitehead Scholars

Duke Center for Statistical Genetics and Genomics
Department of Biostatistics & Bioinformatics
Duke University School of Medicine

DrAlexOchoa@genomic.social
ochoalab.github.io
alejandro.ochoa@duke.edu

New kinship/GRM estimator

Kinship model for neutral genotypes $x_{ij} \in \{0, 1, 2\}$:

$$\mathrm{E}[x_{ij}] = 2p_i, \qquad \mathrm{Cov}(x_{ij}, x_{ik}) = 4p_i\,(1 - p_i)\,\varphi_{jk}.$$

Standard estimator is **biased**:

$$\hat{p}_i = \frac{1}{2n}\sum_{j=1}^{n} x_{ij}, \quad \hat{\varphi}_{jk}^{\mathsf{std}} = \frac{1}{m}\sum_{i=1}^{m} \frac{(x_{ij} - 2\hat{p}_i)\,(x_{ik} - 2\hat{p}_i)}{4\hat{p}_i\,(1 - \hat{p}_i)} \approx \frac{\varphi_{jk} - \bar{\varphi}_j - \bar{\varphi}_k + \bar{\varphi}}{1 - \bar{\varphi}}.$$

`popkin`: first unbiased kinship estimator! R package

$$A_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij} - 1)(x_{ik} - 1) - 1, \qquad \hat{\varphi}_{jk}^{\mathsf{new}} = 1 - \frac{A_{jk}}{\hat{A}_{\mathsf{min}}} \xrightarrow[m\to\infty]{\text{a.s.}} \varphi_{jk}.$$

Ochoa and Storey (2021) doi:10.1371/journal.pgen.1009241

https://github.com/StoreyLab/popkin