Statistical Genetics Research: Kinship, Bias, Admixture

Alejandro Ochoa

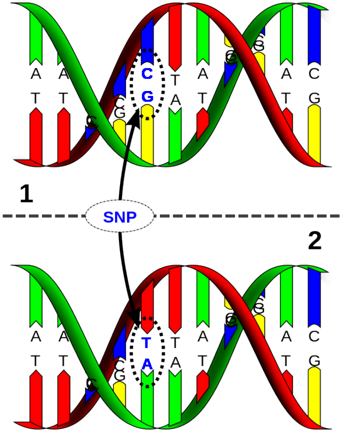
—

🐦 DrAlexOchoa
🏠 ochoalab.github.io
✉ alejandro.ochoa@duke.edu

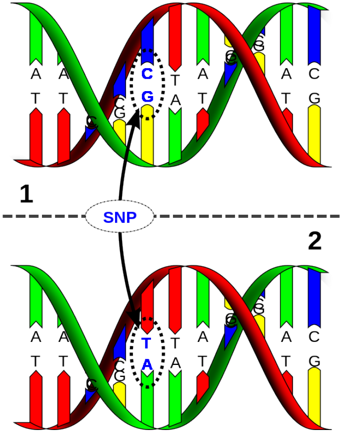
StatGen, Biostatistics & Bioinformatics — Duke University

2021-08-19 — B&B Orientation

Single Nucleotide Polymorphism (SNP) data



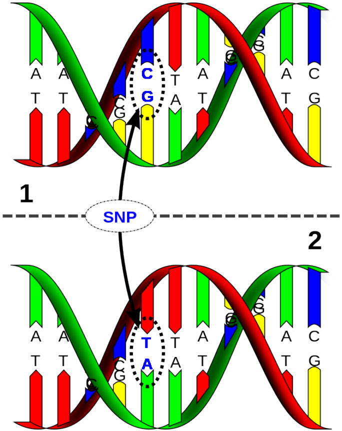
Single Nucleotide Polymorphism (SNP) data



⇒

Genotype	x_{ij}
C/C	0
C/T	1
T/T	2

Single Nucleotide Polymorphism (SNP) data



⇒

Genotype	x_{ij}
C/C	0
C/T	1
T/T	2

⇒

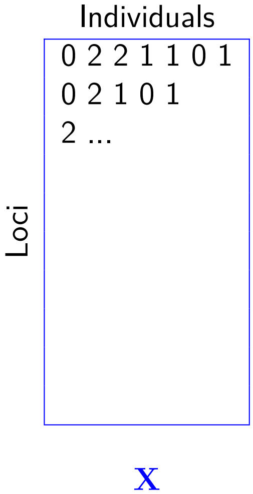
Individuals

Loci

0 2 2 1 1 0 1
0 2 1 0 1
2 ...

$\mathbf{X}$

Dependence structure of genotype matrix



Individuals

0 2 2 1 1 0 1

0 2 1 0 1

2 ...

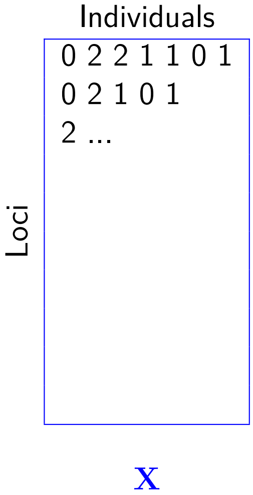
Loci

$\mathbf{X}$

High-dimensional binomial data

- No general likelihood function
- My work: method of moments

Dependence structure of genotype matrix



Individuals

0 2 2 1 1 0 1
0 2 1 0 1
2 ...

Loci

$\mathbf{X}$

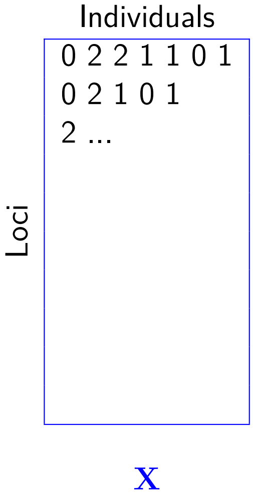
High-dimensional binomial data

- No general likelihood function
- My work: method of moments

Relatedness / Population structure

- Dependence between individuals (columns)

Dependence structure of genotype matrix



Individuals

Loci

0 2 2 1 1 0 1

0 2 1 0 1

2 ...

$\mathbf{X}$

High-dimensional binomial data

- No general likelihood function
- My work: method of moments

Relatedness / Population structure

- Dependence between individuals (columns)

Linkage disequilibrium

- Dependence between loci (rows)

New kinship estimator for general relatedness

New kinship estimator for general relatedness

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

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

New kinship estimator for general relatedness

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**:

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

New kinship estimator for general relatedness

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

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

Standard estimator is **biased**:

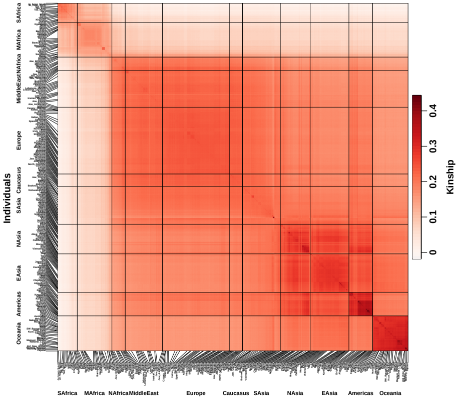
$$\hat{p}_i = \frac{1}{2n}\sum_{j=1}^{n} x_{ij}, \quad \hat{\varphi}_{jk}^{\text{std}} = \frac{1}{m}\sum_{i=1}^{m} \frac{\left(x_{ij} - 2\hat{p}_i\right)\left(x_{ik} - 2\hat{p}_i\right)}{4\hat{p}_i\left(1 - \hat{p}_i\right)} \approx \frac{\varphi_{jk} - \bar{\varphi}_j - \bar{\varphi}_k + \bar{\varphi}}{1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

`popkin`: first unbiased kinship estimator! R package (Ochoa and Storey, 2021)

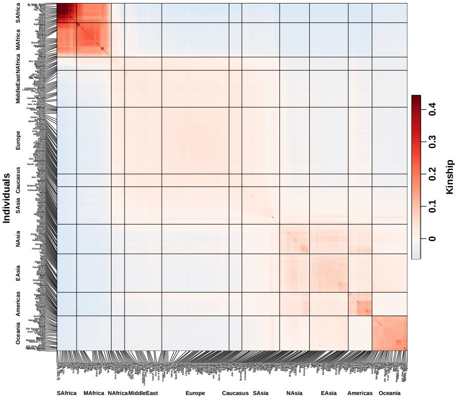
$$A_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij} - 1)(x_{ik} - 1) - 1, \qquad \hat{\varphi}_{jk}^{\text{new}} = 1 - \frac{A_{jk}}{\hat{A}_{\text{min}}} \xrightarrow[m \to \infty]{\text{a.s.}} \varphi_{jk}.$$

https://github.com/StoreyLab/popkin

Kinship bias: Consequences? Applications?



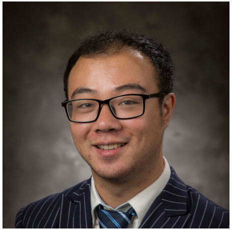
New "popkin"
kinship estimator



Biased "standard"
kinship estimator

Ochoa and Storey (2019b) doi:10.1101/653279

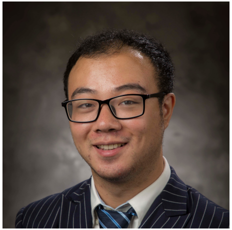
Principal components vs mixed effects in genetic association



Yiqi Yao
MB 2020

BenHealth
Shanghai

Principal components vs mixed effects in genetic association

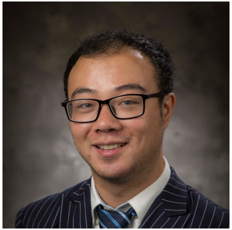


Yiqi Yao
MB 2020

BenHealth
Shanghai

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

Principal components vs mixed effects in genetic association



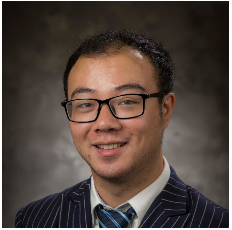
Yiqi Yao
MB 2020

BenHealth
Shanghai

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

$$\begin{aligned} \text{PCA}: \quad & \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{U}_d\boldsymbol{\gamma}_d + \epsilon, \\ \text{LMM}: \quad & \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{s} + \epsilon. \end{aligned}$$

Principal components vs mixed effects in genetic association



Yiqi Yao
MB 2020

BenHealth
Shanghai

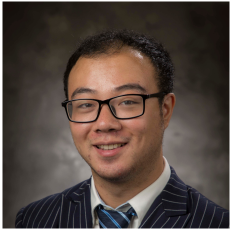
Association with Principal Components Analysis (PCA) and Linear Mixed-effects Model (LMM):

$$\begin{aligned} \text{PCA :} \quad & \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{U}_d\boldsymbol{\gamma}_d + \epsilon, \\ \text{LMM :} \quad & \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{s} + \epsilon. \end{aligned}$$

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

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

Principal components vs mixed effects in genetic association



Yiqi Yao
MB 2020

BenHealth
Shanghai

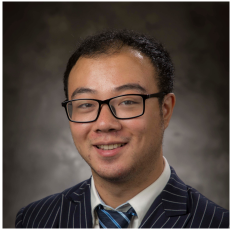
Association with Principal Components Analysis (PCA) and Linear Mixed-effects Model (LMM):

$$\begin{aligned} \text{PCA}: \quad & \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{U}_d\boldsymbol{\gamma}_d + \boldsymbol{\epsilon}, \\ \text{LMM}: \quad & \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{s} + \boldsymbol{\epsilon}. \end{aligned}$$

$\mathbf{U}_d$ are top d eigenvectors of kinship matrix Φ.
$\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, \sigma^2\Phi\right)$.

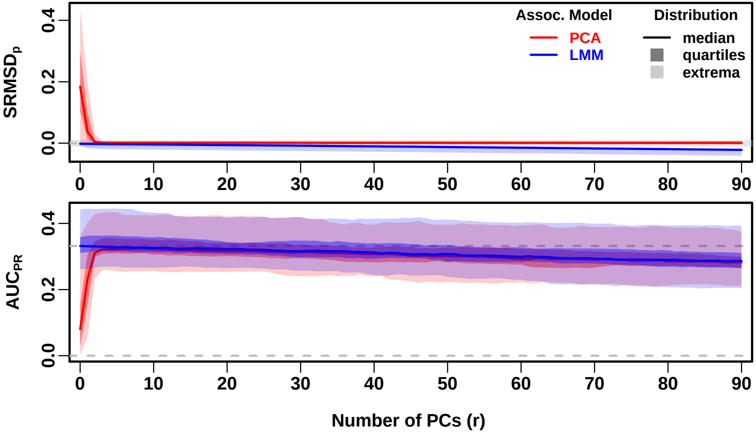
- PCA is faster but low-dimensional
- LMM is slower but can model families
- Both depend on estimated kinship matrix

Principal components vs mixed effects in genetic association



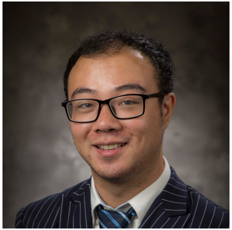
Yiqi Yao
MB 2020

BenHealth
Shanghai



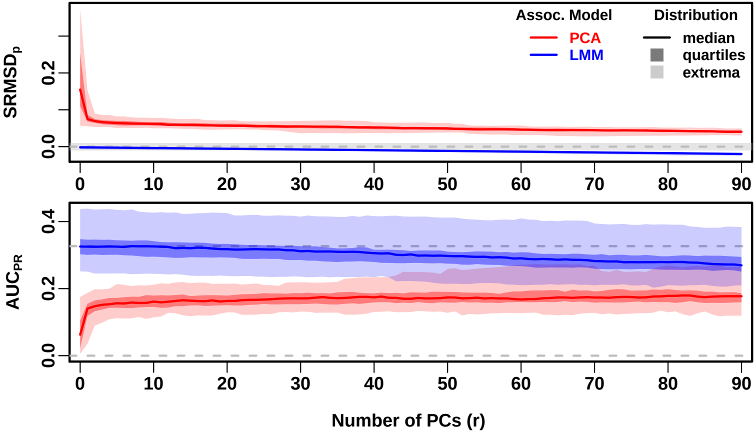
Simulated admixed individuals

Principal components vs mixed effects in genetic association



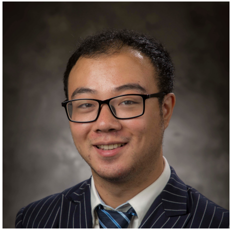
Yiqi Yao
MB 2020

BenHealth
Shanghai



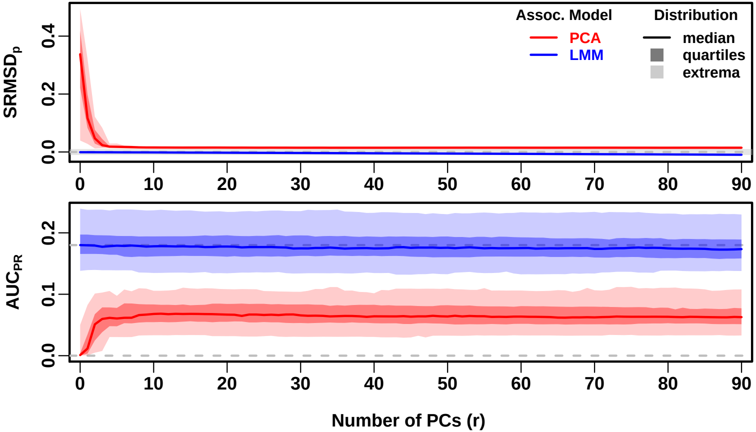
Simulated admixed family

Principal components vs mixed effects in genetic association



Yiqi Yao
MB 2020

BenHealth
Shanghai



1000 Genomes Project

Kinship bias does not affect genetic associations

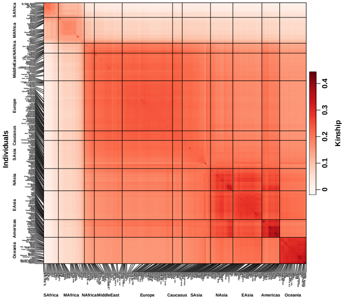
Zhuoran Hou
MB 2021

Now: B&B PhD

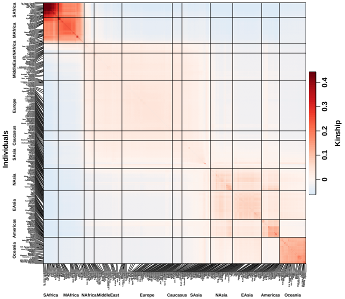
Kinship bias does not affect genetic associations

Zhuoran Hou
MB 2021

Now: B&B PhD



New popkin
kinship estimator

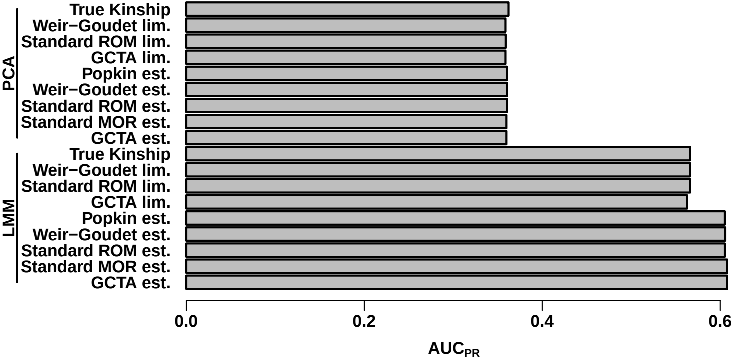


Standard
kinship estimator

Kinship bias does not affect genetic associations

Zhuoran Hou
MB 2021

Now: B&B PhD



Kinship bias doesn't matter?

Kinship bias does not affect genetic associations

Zhuoran Hou
MB 2021

Now: B&B PhD

Proved with linear algebra!

$$\Phi' = \frac{1}{1-\bar{\varphi}} \mathbf{C} \Phi \mathbf{C}, \qquad \mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{J}.$$

Kinship bias does not affect genetic associations

Zhuoran Hou
MB 2021

Now: B&B PhD

Proved with linear algebra!

$$\Phi' = \frac{1}{1 - \bar{\varphi}} \mathbf{C} \Phi \mathbf{C}, \qquad \mathbf{C} = \mathbf{I} - \frac{1}{n} \mathbf{J}.$$

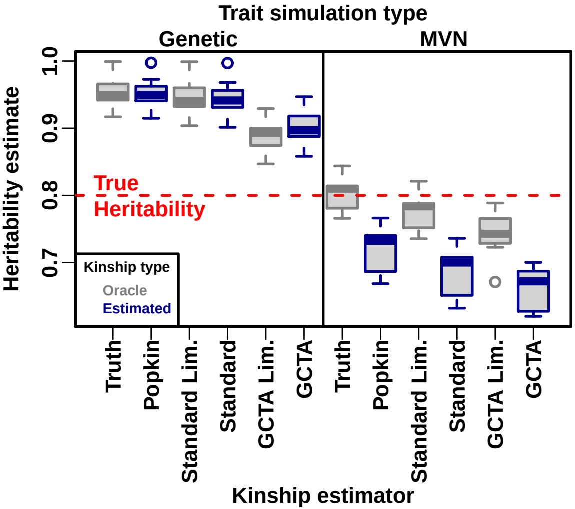
In LMM and PCA, the bias is compensated by the scale and intercept coefficients:

$$\sigma' = \sigma \sqrt{1 - \bar{\varphi}},$$
$$\alpha' = \alpha + \sigma \frac{1}{n} \mathbf{1}^\top \Phi^{\frac{1}{2}} \mathbf{r}.$$

Kinship bias affects heritability estimation

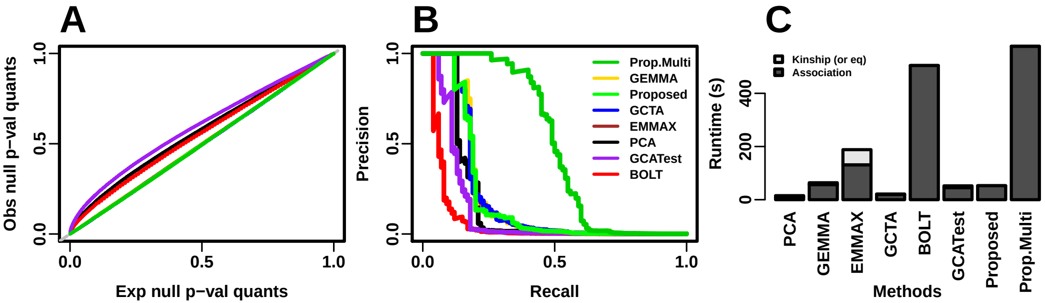
Zhuoran Hou
MB 2021

Now: B&B PhD

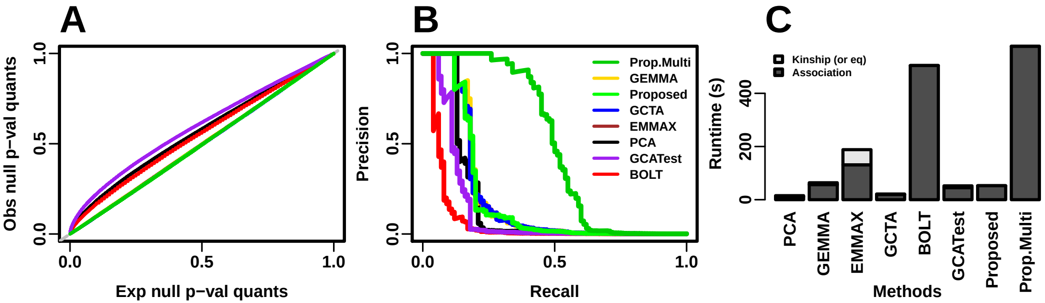


LIGERA: light genetic robust association

LIGERA: light genetic robust association

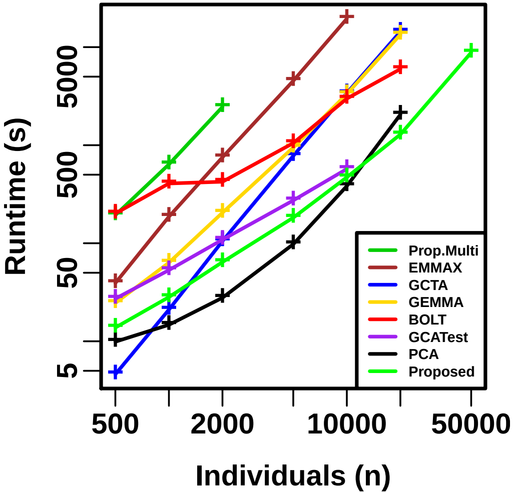


LIGERA: light genetic robust association



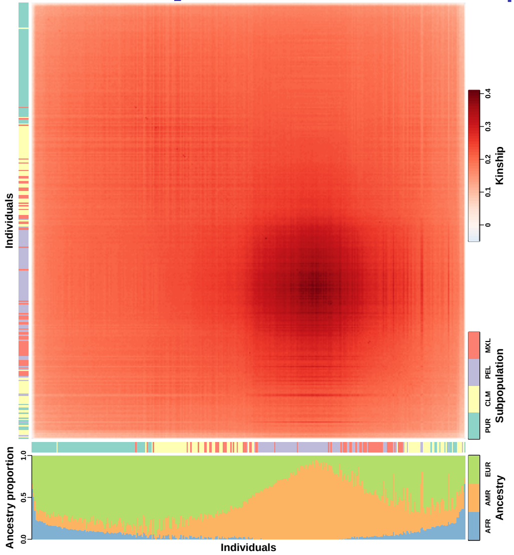
- Control of type-I error
- Increased power with multiscan
- Great runtime for single scan (enables multiscan)

LIGERA: light genetic robust association: scalability



Admixture: kinship driven by admixture in Hispanics

Admixture: kinship driven by admixture in Hispanics



Ochoa and Storey (2019b) doi:10.1101/653279

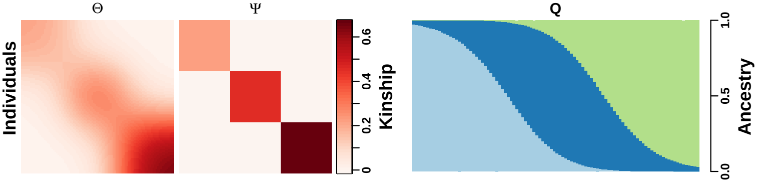
https://github.com/StoreyLab/popkin

Kinship under the admixture model

Amika Sood
Postdoc

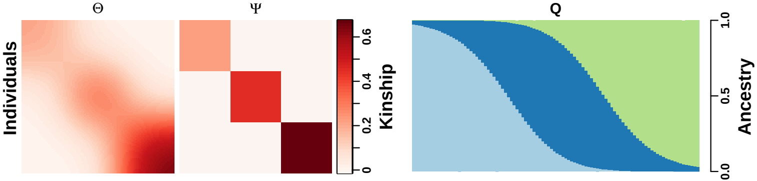
Kinship under the admixture model

Amika Sood
Postdoc



Kinship under the admixture model

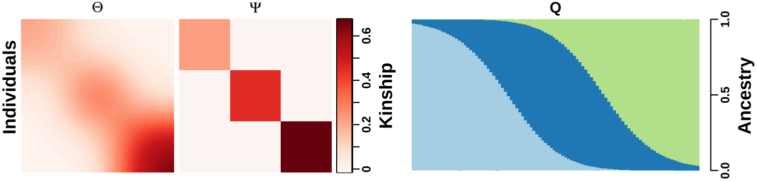
Amika Sood
Postdoc



$$\Theta = \mathbf{Q}\Psi\mathbf{Q}^\top$$

Kinship under the admixture model

Amika Sood
Postdoc

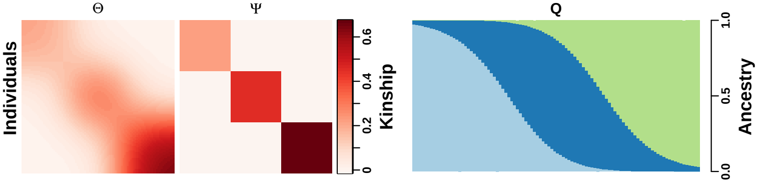


$$\Theta = \mathbf{Q}\Psi\mathbf{Q}^\top$$

Can we reverse this formula?

Kinship under the admixture model

Amika Sood
Postdoc



$$\Theta = \mathbf{Q}\Psi\mathbf{Q}^\top$$

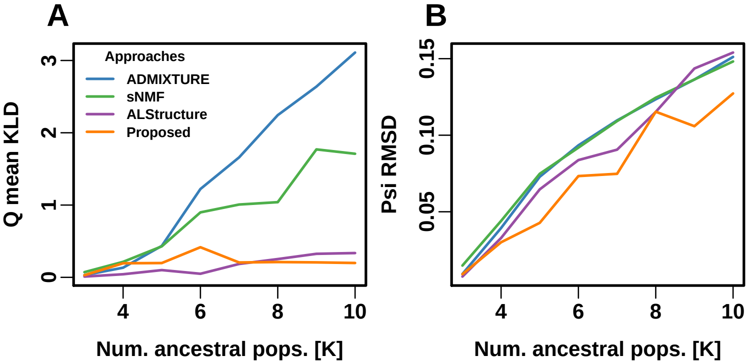
Can we reverse this formula?

Constrained optimization, regularized objective:

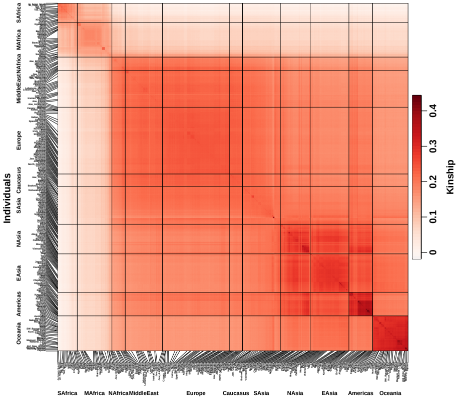
$$F = \left|\left|\hat{\Theta} - \mathbf{Q}\Psi\mathbf{Q}^\top\right|\right|^2 + \gamma\,\mathrm{tr}(\Psi).$$

AdmixCor: accuracy

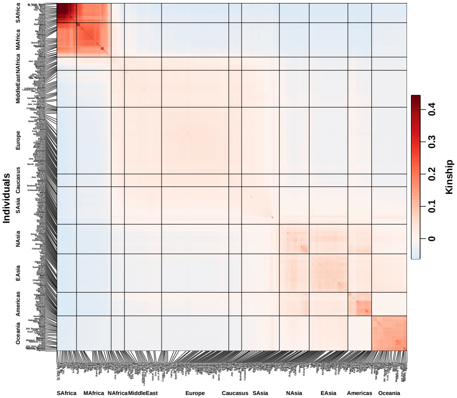
Amika Sood
Postdoc



Unbiased kinship estimates: new models, opportunities



New ”popkin”
kinship estimator



Biased ”standard”
kinship estimator

Ochoa and Storey (2019b) doi:10.1101/653279

Acknowledgments

Ochoa Lab

Amika Sood
Tiffany Tu
Yiqi Yao
Zhuoran Hou
Jiajie Shen
Emmanuel Mokel

Princeton University

John D. Storey

Duke University

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

Funding

NIH
Whitehead Scholars

Duke Center for Statistical Genetics and Genomics

GCB
Duke Center for Genomic and Computational Biology

Department of Biostatistics & Bioinformatics
Duke University School of Medicine

🐦 DrAlexOchoa
🏠 ochoalab.github.io

✉ alejandro.ochoa@duke.edu