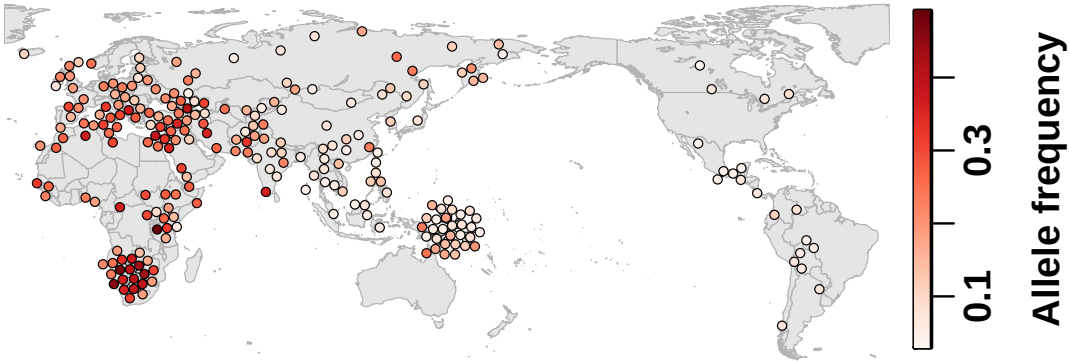
Statistical Genetics Research: Kinship, Bias, Admixture

Alejandro Ochoa

StatGen, Biostatistics & Bioinformatics — Duke University

2022-06-11 — UPGG retreat

Human genetic structure

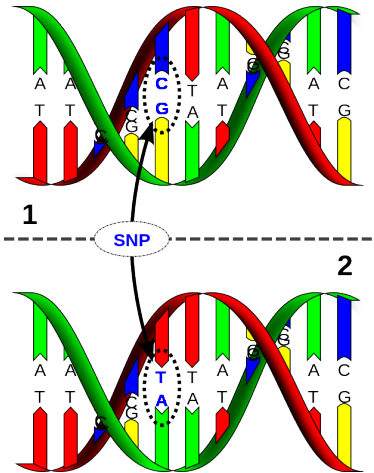


Ochoa and Storey (2019a) doi:10.1101/653279

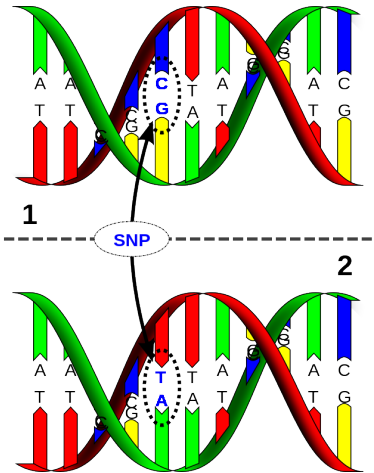
rs17110306; median differentiation given MAF $\geq 10\%$

Why? Migration and isolation, admixture, family structure

Single Nucleotide Polymorphism (SNP) data



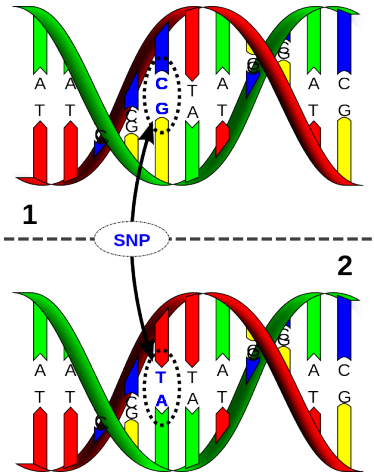
Single Nucleotide Polymorphism (SNP) data



$\Rightarrow$

Genotype	x_{ij}
CC	0
CT	1
TT	2

Single Nucleotide Polymorphism (SNP) data



$\Rightarrow$

Genotype	x_{ij}
CC	0
CT	1
TT	2

$\Rightarrow$

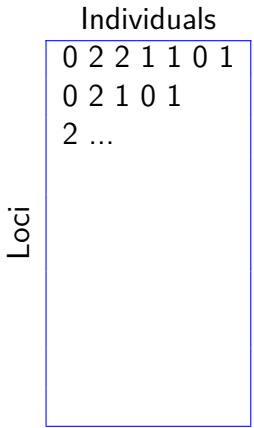
Individuals

Loci

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

$\mathbf{X}$

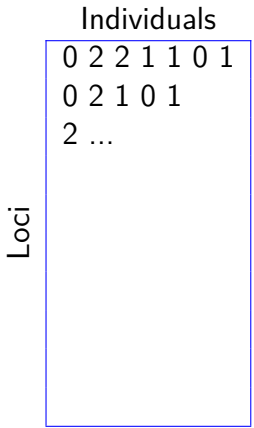
Dependence structure of genotype matrix



High-dimensional binomial data

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

Dependence structure of genotype matrix



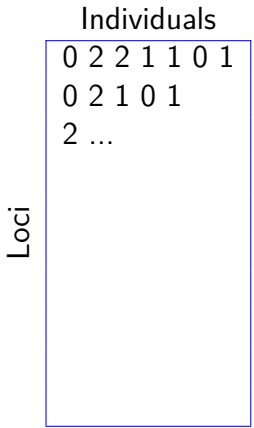
High-dimensional binomial data

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

Relatedness / Population structure

- Dependence between individuals (columns)

Dependence structure of genotype matrix



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/GRM estimator for general relatedness

New kinship/GRM estimator for general relatedness

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

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

New kinship/GRM 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**:

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

New kinship/GRM 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**:

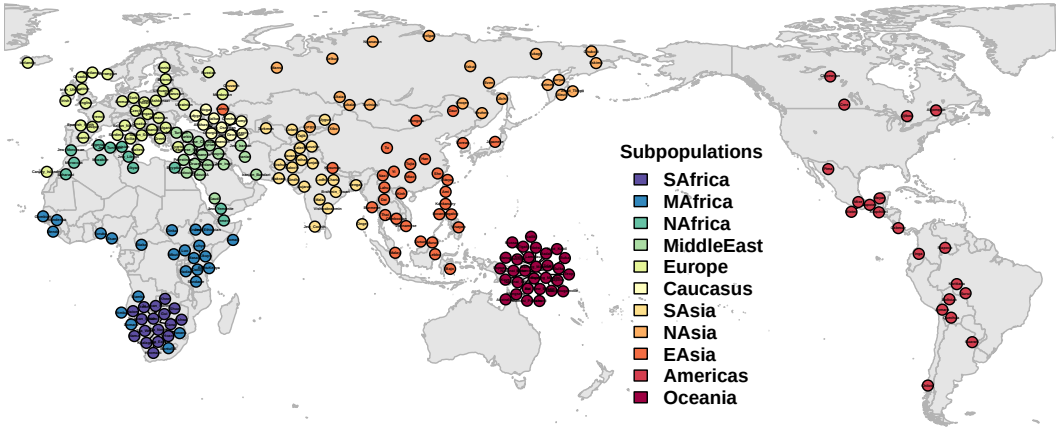
$$\hat{p}_i = \frac{1}{2n}\sum_{j=1}^{n} x_{ij}, \quad \hat{\varphi}_{jk}^{\text{std}} = \frac{1}{m}\sum_{i=1}^{m} \frac{(x_{ij} - 2\hat{p}_i)\,(x_{ik} - 2\hat{p}_i)}{4\hat{p}_i\,(1 - \hat{p}_i)} \approx \frac{\varphi_{jk} - \bar{\varphi}_j - \bar{\varphi}_k + \bar{\varphi}}{1 - \bar{\varphi}}.$$

popkin: first unbiased kinship estimator! R package (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

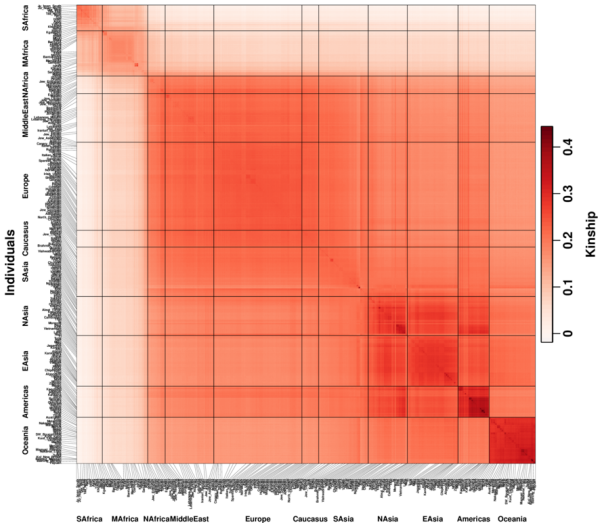
Dataset: Human Origins

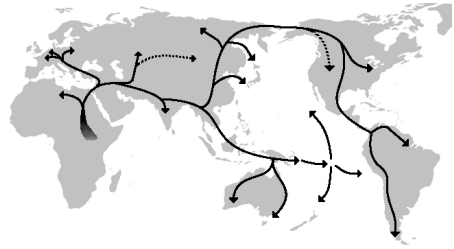


Lazaridis *et al.* (2014), (2016); Skoglund *et al.* (2016)

2,922 indivs. from 243 locs. — 588,091 loci — Array

Kinship matrix of world-wide human population

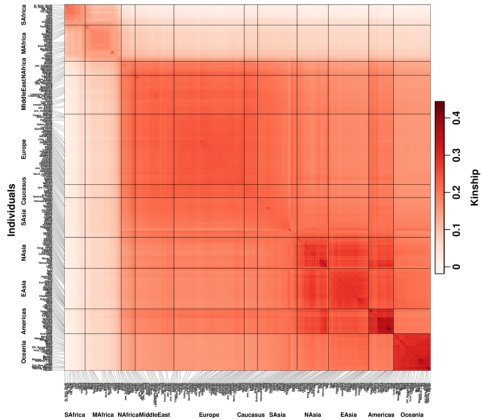




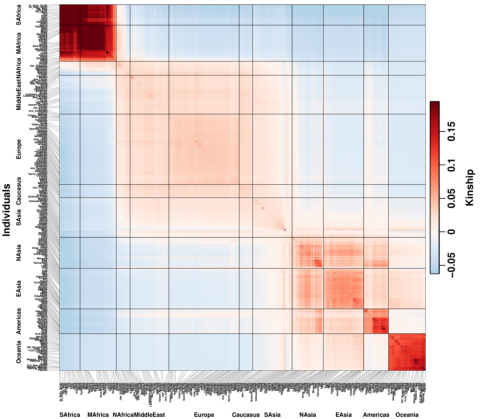
Ochoa and Storey (2019) doi:10.1101/653279

Standard kinship estimator is severely biased

New



Standard

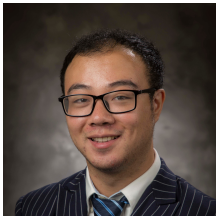


Ochoa and Storey (2019) doi:10.1101/653279

Kinship bias: Consequences? Applications?

- Genetic association studies
- Heritability estimation
- Admixture inference

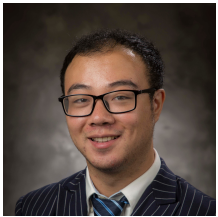
PCA vs LMM in association



Yiqi Yao
MB 2020

BenHealth
Shanghai

PCA vs LMM in association

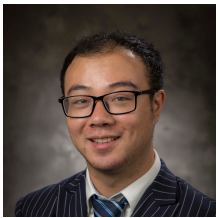


Yiqi Yao
MB 2020

BenHealth
Shanghai

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

PCA vs LMM in association



Yiqi Yao
MB 2020

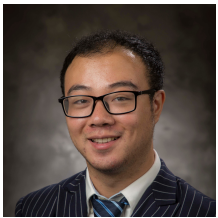
BenHealth
Shanghai

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

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

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

PCA vs LMM in association



Yiqi Yao
MB 2020

BenHealth
Shanghai

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

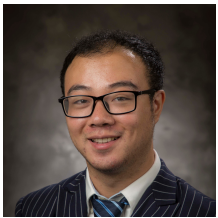
$$\text{PCA}: \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{U}_d\gamma_d + \epsilon,$$

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

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

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

PCA vs LMM in association



Yiqi Yao
MB 2020

BenHealth
Shanghai

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

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

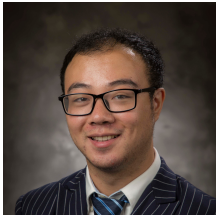
$$\text{LMM}: \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{s} + \epsilon.$$

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

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

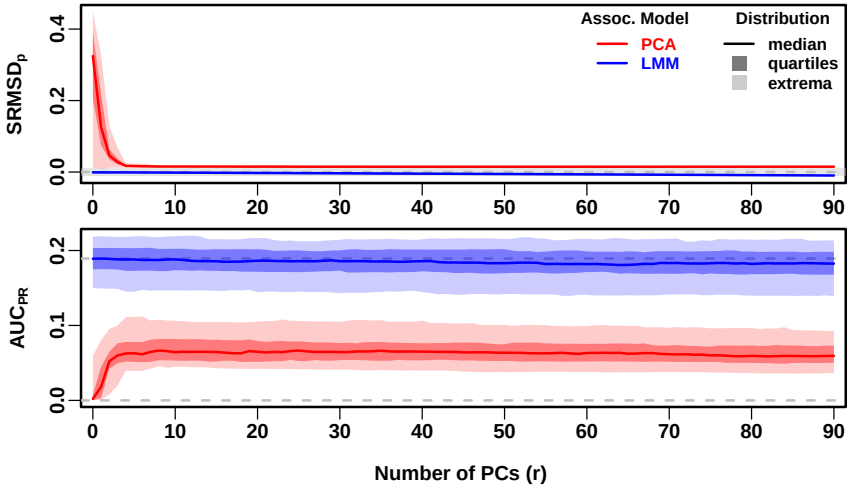
Yao and Ochoa (2022) doi:10.1101/2022.03.25.485885

PCA < LMM in association for real datasets



Yiqi Yao
MB 2020

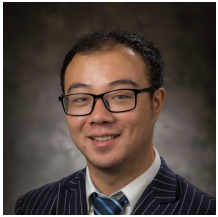
BenHealth
Shanghai



1000 Genomes Project

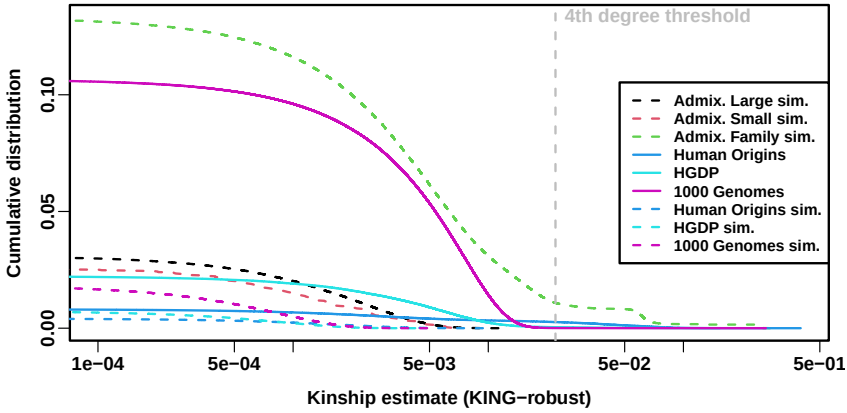
Yao and Ochoa (2022) doi:10.1101/2022.03.25.485885

Numerous distant relatives in real datasets



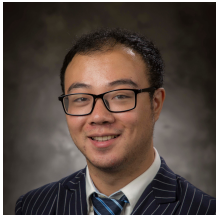
Yiqi Yao
MB 2020

BenHealth
Shanghai



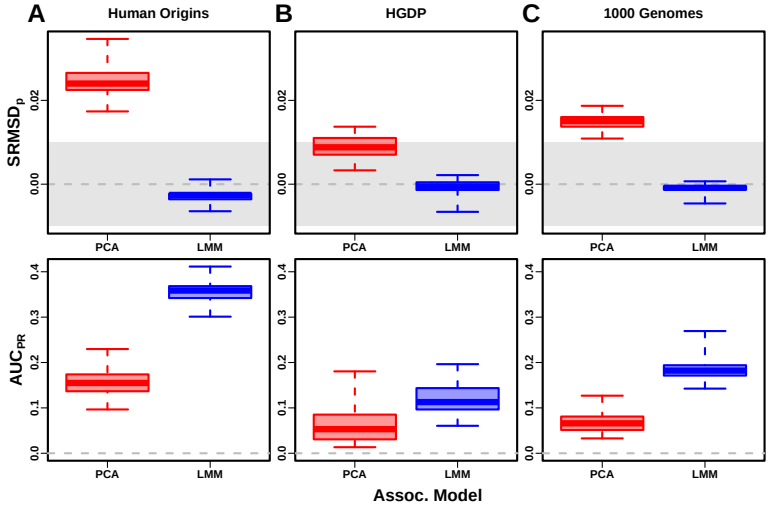
Yao and Ochoa (2022) doi:10.1101/2022.03.25.485885

Numerous distant relatives in real datasets explain PCA < LMM



Yiqi Yao
MB 2020

BenHealth
Shanghai

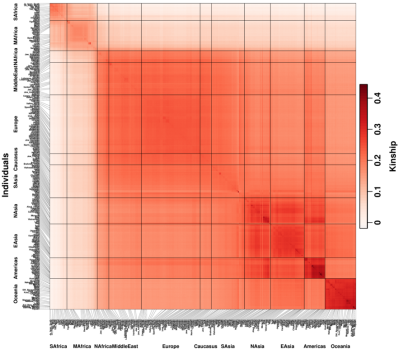


Yao and Ochoa (2022) doi:10.1101/2022.03.25.485885

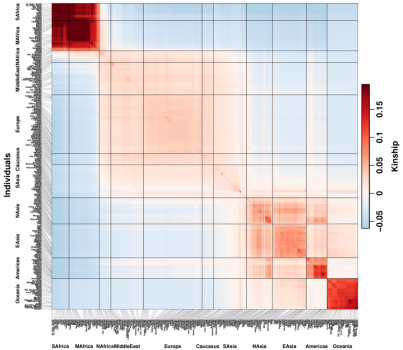
Kinship bias does not affect genetic associations

Zhuoran Hou
MB 2021

Now: B&B PhD student



New popkin kinship estimator

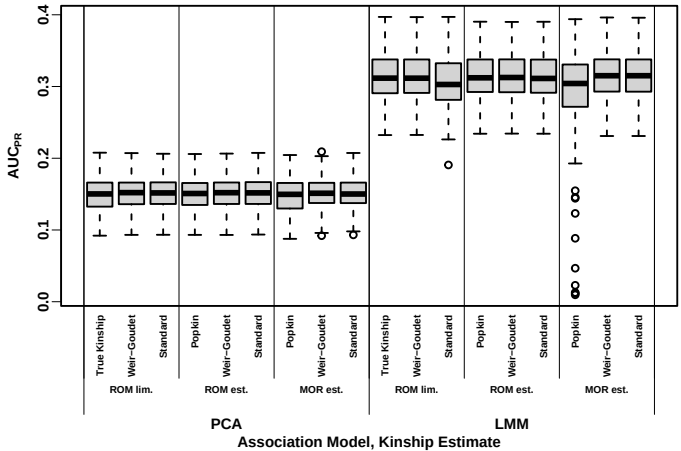


Standard kinship estimator

Kinship bias does not affect genetic associations

Zhuoran Hou
MB 2021

Now: B&B PhD student

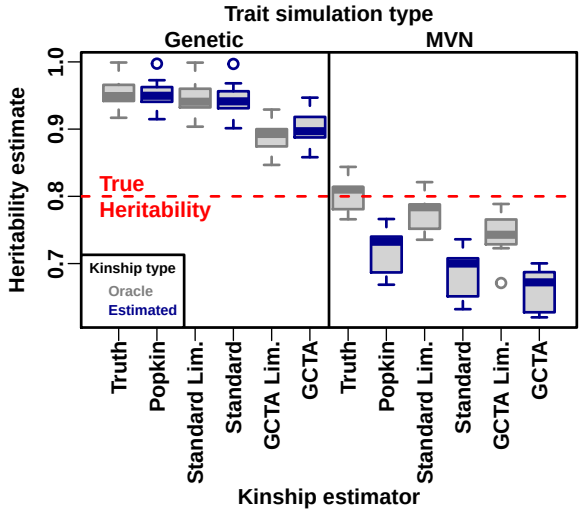


Kinship bias doesn't matter? Proved with linear algebra!

Kinship bias affects heritability estimation

Zhuoran Hou
MB 2021

Now: B&B PhD student



Nephrotic Syndrome association study

Tiffany Tu
CBB PhD student

With
Gbadegesin lab
(Pediatrics)

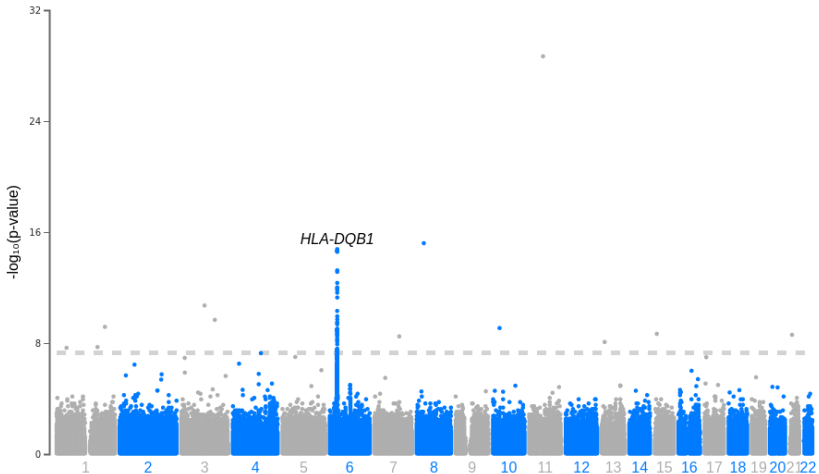
Severe pediatric kidney disease. 1000 cases/1000 controls; multiethnic; LMM association

Nephrotic Syndrome association study

Tiffany Tu
CBB PhD student

With Gbadegesin lab (Pediatrics)

Severe pediatric kidney disease. 1000 cases/1000 controls; multiethnic; LMM association



LIGERA: light genetic robust association

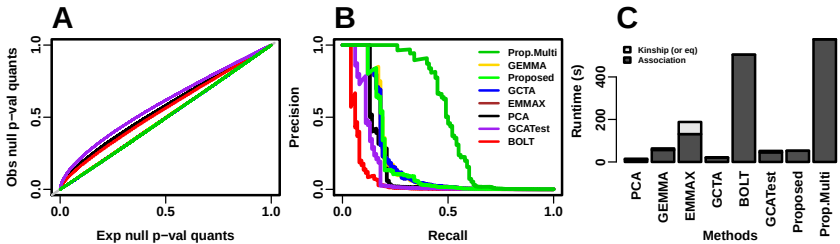
Tiffany Tu
CBB PhD
student

- Enabled by unbiased kinship
- Like LMM but reverses trait and genotype
  - Faster by avoiding costly “variant component” fitting

LIGERA: light genetic robust association

Tiffany Tu
CBB PhD
student

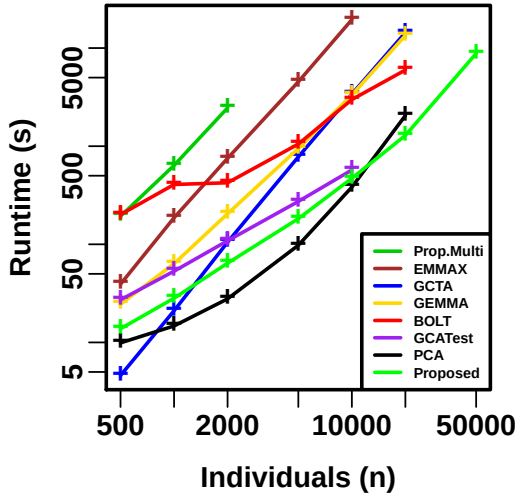
- Enabled by unbiased kinship
- Like LMM but reverses trait and genotype
  - Faster by avoiding costly "variant component" fitting



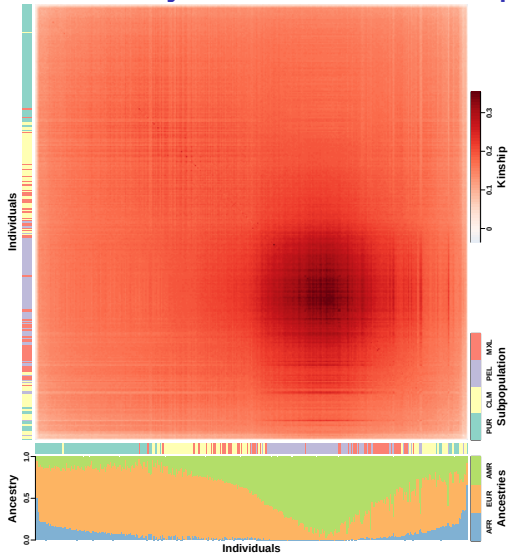
- Control of type-I error
- Increased power with multiscan: polygenic model!

LIGERA: light genetic robust association: scalability

Tiffany Tu
CBB PhD
student



Population kinship driven by admixture in Hispanics



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

Kinship under the admixture model

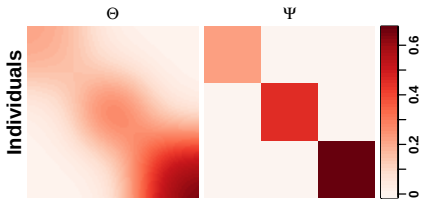
Amika Sood
Postdoc

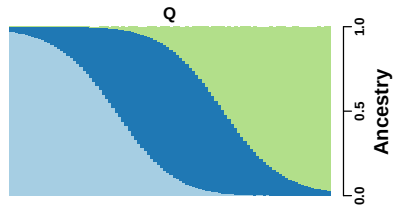
Now: Senior Researcher
UGA

Kinship under the admixture model

Amika Sood
Postdoc

Now: Senior Researcher
UGA

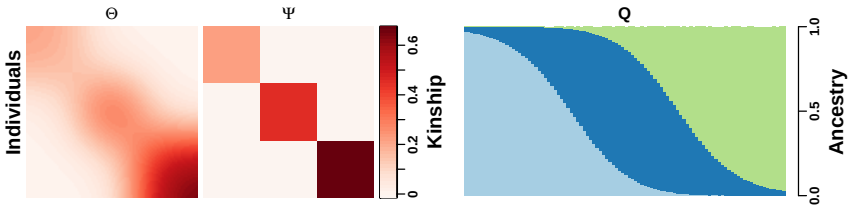




Kinship under the admixture model

Amika Sood
Postdoc

Now: Senior Researcher
UGA



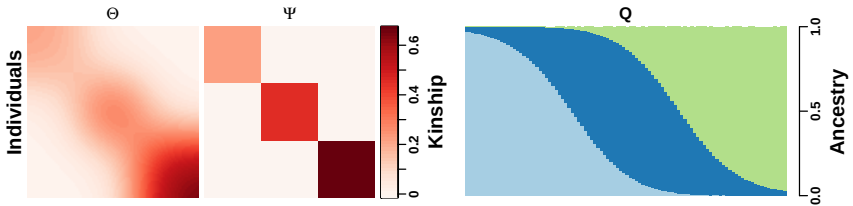
$$\Theta = \mathbf{Q}\Psi\mathbf{Q}^\top$$

(Only for unbiased kinship)

Kinship under the admixture model

Amika Sood
Postdoc

Now: Senior Researcher
UGA



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

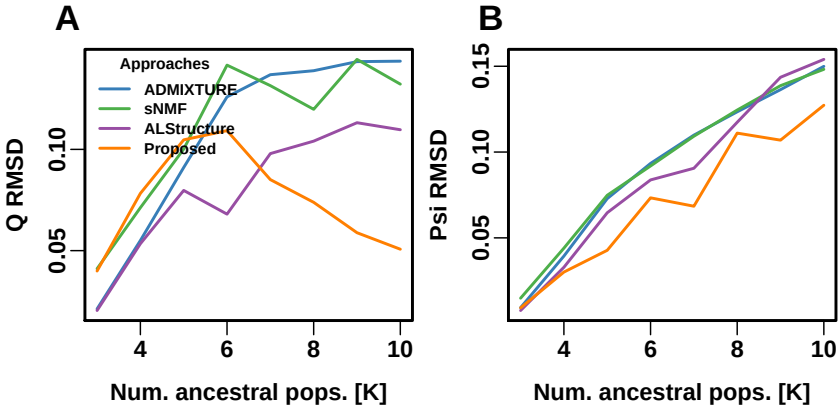
(Only for unbiased kinship)

Can we reverse this formula?

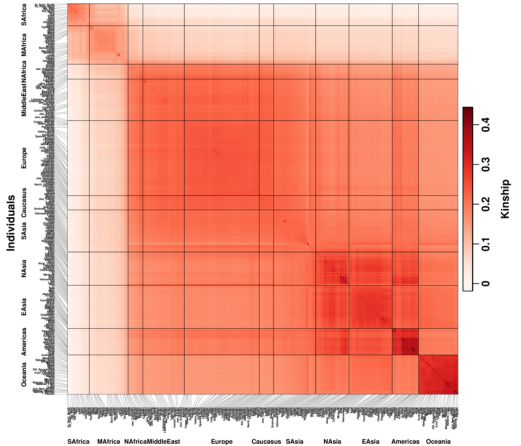
AdmixCor: accuracy

Amika Sood
Postdoc

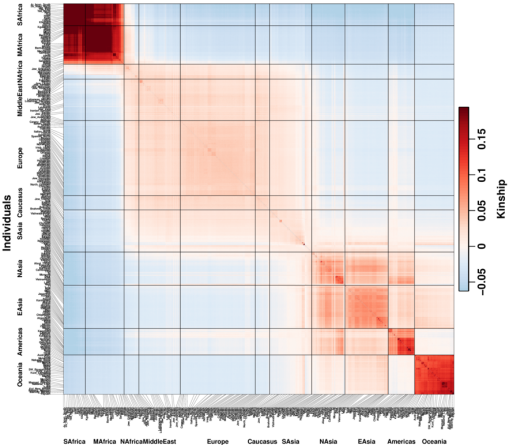
Now: Senior Researcher
UGA



Unbiased kinship estimates: new models, opportunities



New "popkin"
kinship estimator



Biased "standard"
kinship estimator

Acknowledgments

Ochoa Lab
Amika Sood
Tiffany Tu
Yiqi Yao
Zhuoran Hou
Jiajie Shen
Emmanuel Mokel
RP Pornmongkolsuk

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
ochoalab.github.io
alejandro.ochoa@duke.edu