

Statistical Genetics Research: Kinship, Bias, Admixture

Alejandro Ochoa

StatGen, Biostatistics & Bioinformatics — Duke University



2022-07-08 — LatMath conference, UCLA IPAM

How did I get here? A zigzaggy line



How did I get here? A zigzaggy line

- ▶ Born in El Paso, Texas , grew up in Ciudad Juárez, México 




How did I get here? A zigzaggy line

- ▶ Born in El Paso, Texas , grew up in Ciudad Juárez, México 
- ▶ High school: math olympiad in Mexico
 - ▶ News: Human Genome Project (2000)

How did I get here? A zigzaggy line

- ▶ Born in El Paso, Texas , grew up in Ciudad Juárez, México 
- ▶ High school: math olympiad in Mexico
 - ▶ News: Human Genome Project (2000)
- ▶ College: MIT
 - ▶ Started Bio major; added Math minor; ended up Bio + Math double major
 - ▶ No applied math/stats 😬
 - ▶ UG research: computational protein design

How did I get here? A zigzaggy line

- ▶ Born in El Paso, Texas , grew up in Ciudad Juárez, México 
- ▶ High school: math olympiad in Mexico
 - ▶ News: Human Genome Project (2000)
- ▶ College: MIT
 - ▶ Started Bio major; added Math minor; ended up Bio + Math double major
 - ▶ No applied math/stats 😱
 - ▶ UG research: computational protein design
- ▶ PhD: Princeton, Molecular Bio (really Computational Bio)
 - ▶ First serious exposure to p-values, permutation tests
 - ▶ Joined a Comp Bio lab: protein domain prediction + malaria 
 - ▶ Late started doing real statistics: q-values, IFDRs

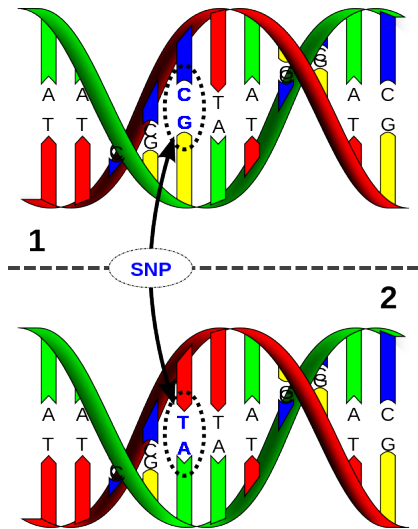
How did I get here? A zigzaggy line

- ▶ Born in El Paso, Texas 🇺🇸, grew up in Ciudad Juárez, México 🇲🇪
- ▶ High school: math olympiad in Mexico
 - ▶ News: Human Genome Project (2000)
- ▶ College: MIT
 - ▶ Started Bio major; added Math minor; ended up Bio + Math double major
 - ▶ No applied math/stats 😱
 - ▶ UG research: computational protein design
- ▶ PhD: Princeton, Molecular Bio (really Computational Bio)
 - ▶ First serious exposure to p-values, permutation tests
 - ▶ Joined a Comp Bio lab: protein domain prediction + malaria 🦟
 - ▶ Late started doing real statistics: q-values, IFDRs
- ▶ Postdoc: Princeton
 - ▶ Switched to human statistical genetics! 🧬
 - ▶ Calculated bias of common estimators, derived new unbiased estimator

How did I get here? A zigzaggy line

- ▶ Born in El Paso, Texas 🇺🇸, grew up in Ciudad Juárez, México 🇲🇪
- ▶ High school: math olympiad in Mexico
 - ▶ News: Human Genome Project (2000)
- ▶ College: MIT
 - ▶ Started Bio major; added Math minor; ended up Bio + Math double major
 - ▶ No applied math/stats 😱
 - ▶ UG research: computational protein design
- ▶ PhD: Princeton, Molecular Bio (really Computational Bio)
 - ▶ First serious exposure to p-values, permutation tests
 - ▶ Joined a Comp Bio lab: protein domain prediction + malaria 🦟
 - ▶ Late started doing real statistics: q-values, IFDRs
- ▶ Postdoc: Princeton
 - ▶ Switched to human statistical genetics! 🧬
 - ▶ Calculated bias of common estimators, derived new unbiased estimator
- ▶ Assistant Professor: Duke, Biostats!

Genetic variation: we're all mutants!



Each newborn has ≈ 70 new mutations!

- ▶ Average mutation rate
 $\approx 1.1 \times 10^{-8}$ /base/generation
 - ▶ Higher in male lineage, with age
- ▶ Number of bases in genome
 $\approx 3.2 \times 10^9$, $\times 2$ for both copies

Dynamics of genetic variation

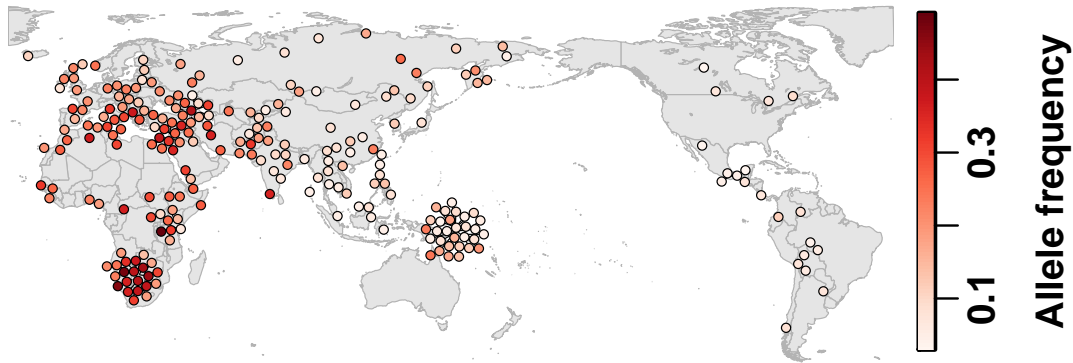


Colors are *alleles*

By Gabi Slizewska

- ▶ Most new mutations are lost
- ▶ Some become common in population
 - ▶ Outcomes are random
 - ▶ Variation greatest in small populations
 - ▶ Even disease alleles can become common

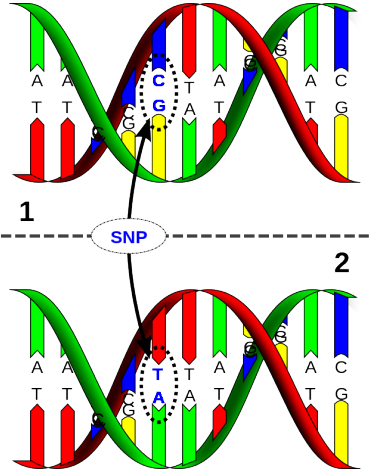
Human genetic structure: a typical allele



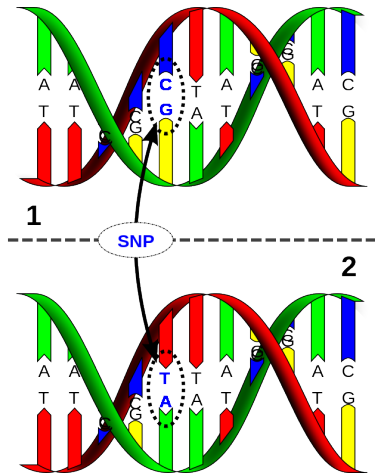
Ochoa and Storey (2019a) doi:10.1101/653279

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

Single Nucleotide Polymorphism (SNP) data

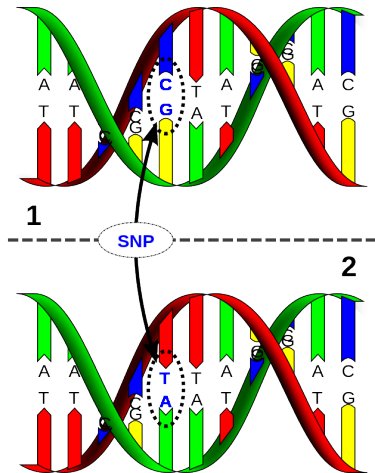


Single Nucleotide Polymorphism (SNP) data


$$\Rightarrow$$

Genotype	x_{ij}
CC	0
CT	1
TT	2

Single Nucleotide Polymorphism (SNP) data



⇒

Genotype	x_{ij}
CC	0
CT	1
TT	2

⇒

	Individuals						
Loci	0	2	2	1	1	0	1
	0	2	1	0	1		
	2	...					
	X						

Dependence structure of genotype matrix

	Individuals						
Loci	0	2	2	1	1	0	1
	0	2	1	0	1		
	2	...					

X

High-dimensional binomial data

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

Dependence structure of genotype matrix

	Individuals						
Loci	0	2	2	1	1	0	1
	0	2	1	0	1		
	2	...					

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	...					

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/GRM estimator

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

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

New kinship/GRM estimator

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

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

Standard estimator is **biased**:

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

New kinship/GRM estimator

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

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

Standard estimator is **biased**:

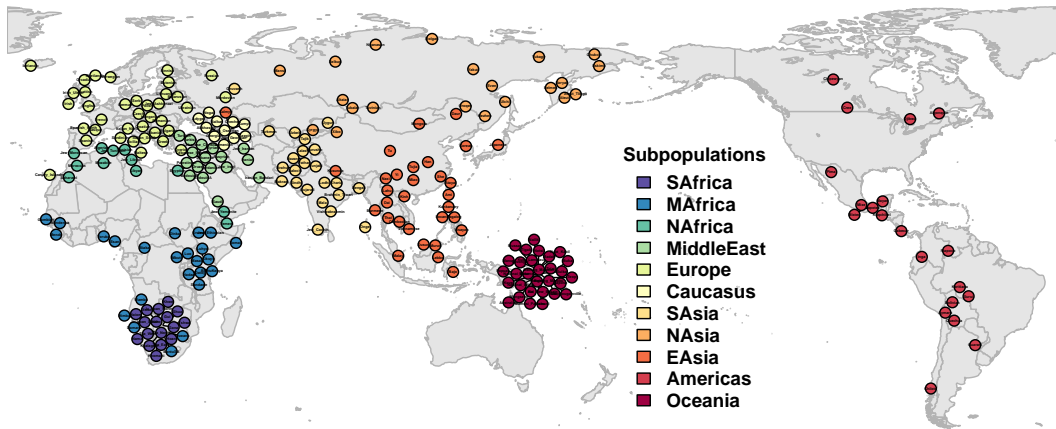
$$\hat{p}_i = \frac{1}{2n} \sum_{j=1}^n x_{ij}, \quad \hat{\varphi}_{jk}^{\text{std}} = \frac{1}{m} \sum_{i=1}^m \frac{(x_{ij} - 2\hat{p}_i)(x_{ik} - 2\hat{p}_i)}{4\hat{p}_i(1 - \hat{p}_i)} \approx \frac{\varphi_{jk} - \bar{\varphi}_j - \bar{\varphi}_k + \bar{\varphi}}{1 - \bar{\varphi}}.$$

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

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



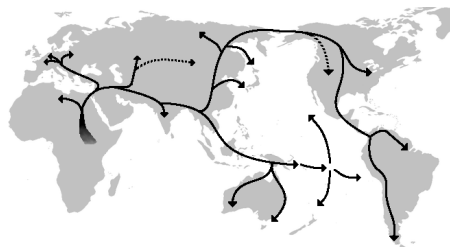
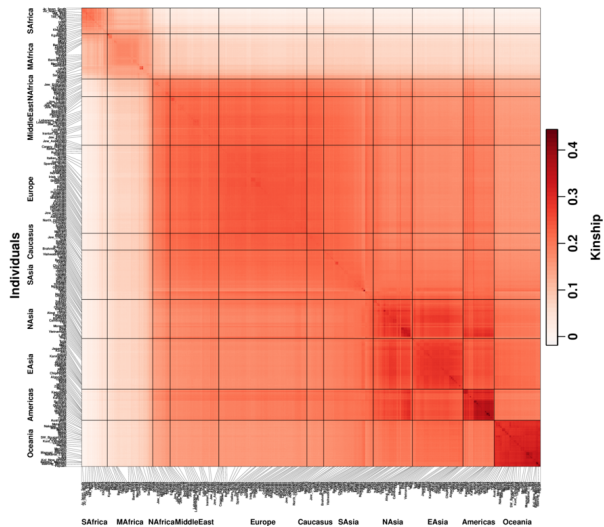
Dataset: Human Origins



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

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

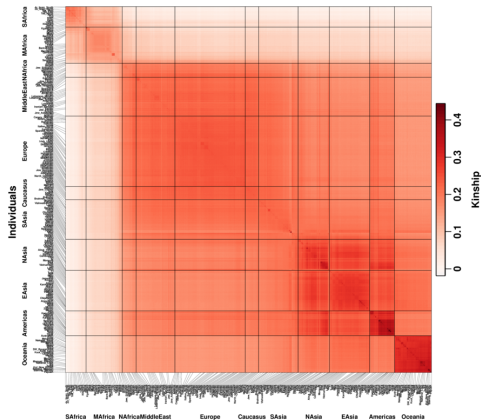
Kinship matrix of world-wide human population



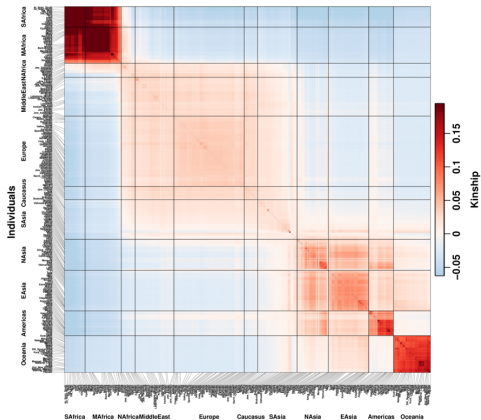
Ochoa and Storey (2019) doi:10.1101/653279

Standard kinship estimator is severely biased

New



Standard



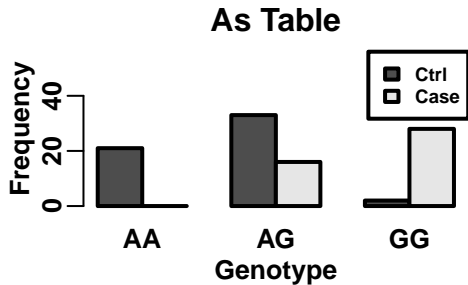
Ochoa and Storey (2019) doi:10.1101/653279

Kinship bias: Consequences? Applications?

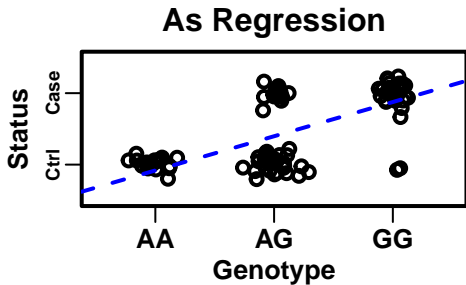
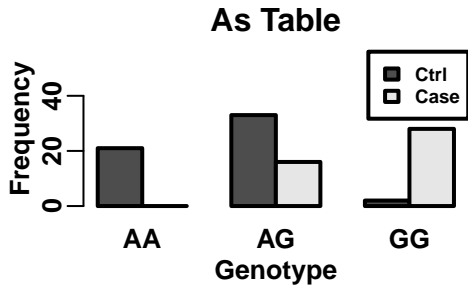
- ▶ Genetic association studies
- ▶ Heritability estimation
- ▶ Admixture inference

Genetic association study: genotype-phenotype correlation

Genetic association study: genotype-phenotype correlation

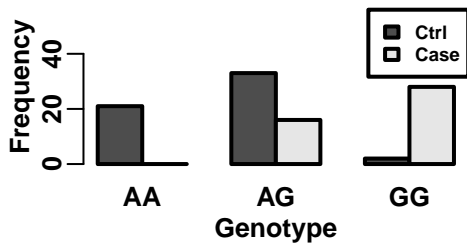


Genetic association study: genotype-phenotype correlation

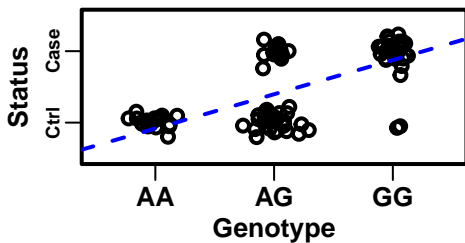


Genetic association study: genotype-phenotype correlation

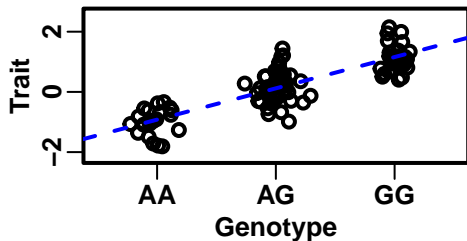
As Table



As Regression

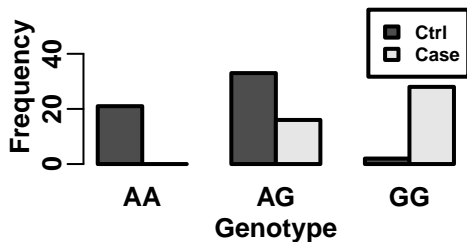


Continuous trait

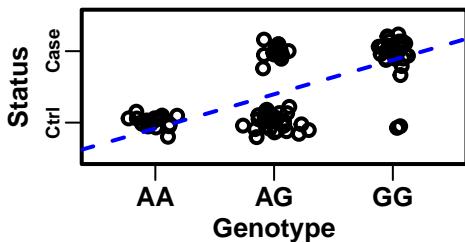


Genetic association study: genotype-phenotype correlation

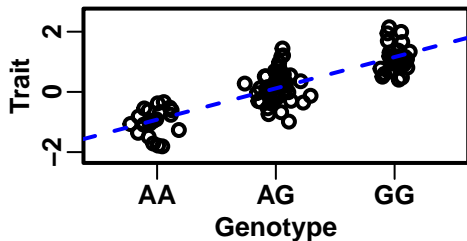
As Table



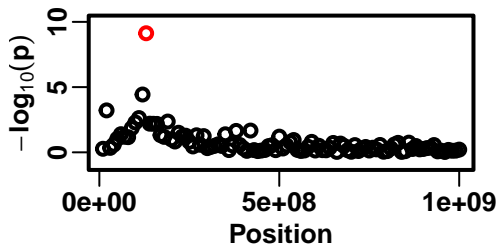
As Regression



Continuous trait

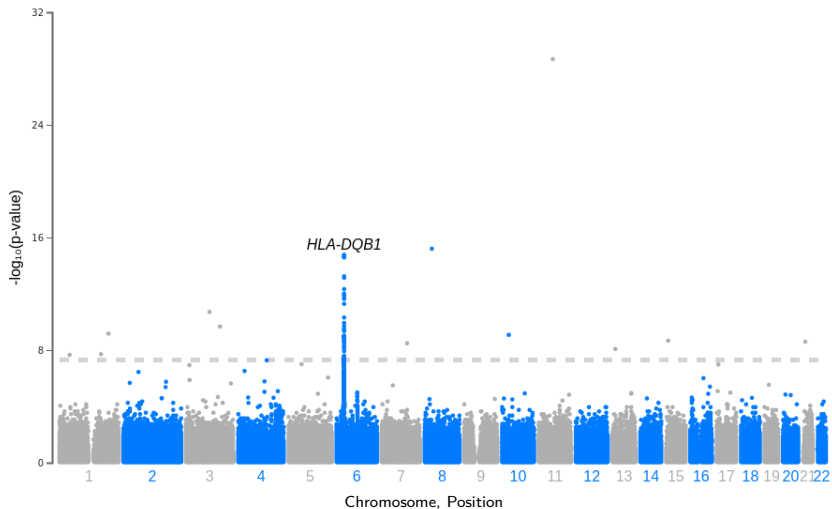


Genome Scan



Nephrotic Syndrome association study

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

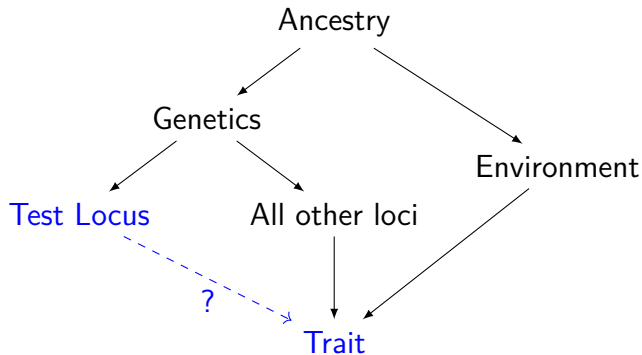


Why is this problem so hard?

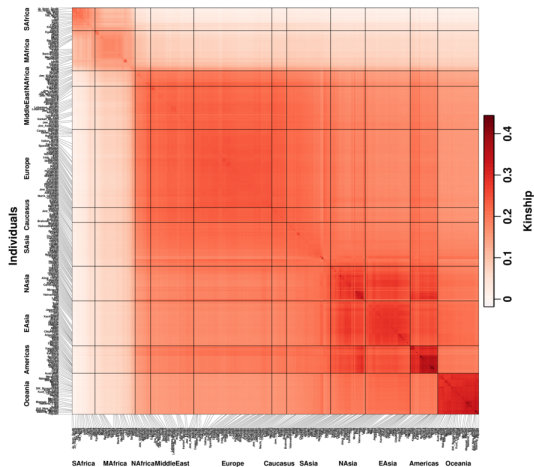
- ▶ Millions of tests
- ▶ Polygenicity (many causal variants)
- ▶ Confounders
- ▶ Incorrect assumptions: independence / additivity

Why is this problem so hard?

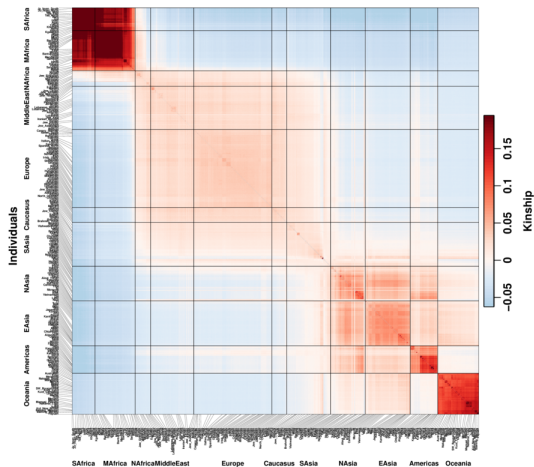
- ▶ Millions of tests
- ▶ Polygenicity (many causal variants)
- ▶ Confounders
- ▶ Incorrect assumptions: independence / additivity



Kinship bias does not affect genetic associations

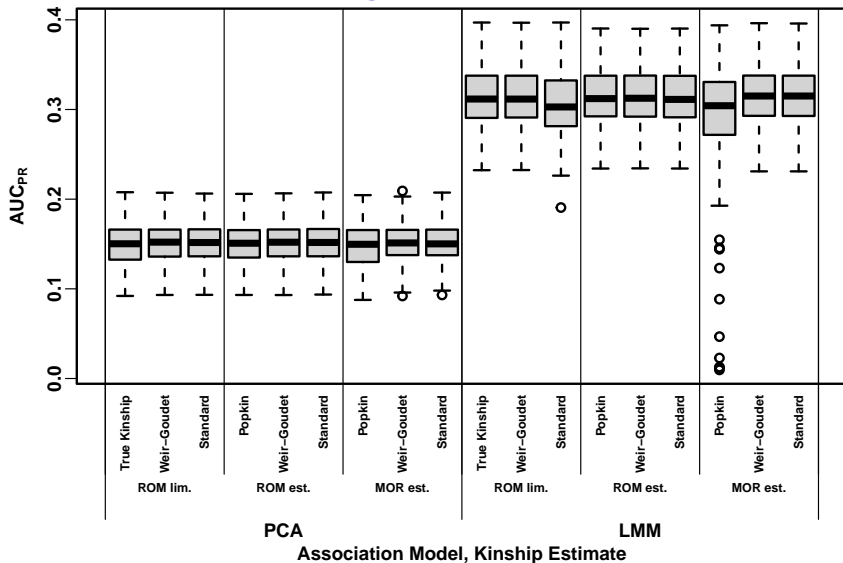


New popkin
kinship estimator



Standard
kinship estimator

Kinship bias does not affect genetic associations



Kinship bias does not affect genetic associations

Linear algebra proof!

Transforming true to biased kinship matrices:

Φ : True kinship matrix,

Φ' : Limit of biased estimator,

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

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

Kinship bias does not affect genetic associations

Linear algebra proof!

Transforming true to biased kinship matrices:

Φ : True kinship matrix,

Φ' : Limit of biased estimator,

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

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

Association test is a regression with correlated residuals:

$$\begin{aligned} \mathbf{y} &= \mathbf{1} \alpha + \mathbf{x}_i \beta_i + \mathbf{s} + \epsilon, \\ \mathbf{s} &\sim \text{Normal}(\mathbf{0}, 2\sigma_G^2 \Phi), \\ \epsilon &\sim \text{Normal}(\mathbf{0}, \sigma_E^2 \mathbf{I}). \end{aligned}$$

Kinship bias does not affect genetic associations

Linear algebra proof!

Transforming true to biased kinship matrices:

Φ : True kinship matrix,

Φ' : Limit of biased estimator,

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

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

Association test is a regression with correlated residuals:

$$\mathbf{y} = \mathbf{1} \alpha + \mathbf{x}_i \beta_i + \mathbf{s} + \epsilon,$$

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

$$\epsilon \sim \text{Normal}(\mathbf{0}, \sigma_E^2 \mathbf{I}).$$

Kinship bias compensated by intercept!

$$\mathbf{s}' = \mathbf{C} \mathbf{s} \sim \text{Normal}(\mathbf{0}, 2\sigma_G'^2 \Phi'),$$

$$\sigma_G'^2 = (1 - \bar{\varphi}) \sigma_G^2,$$

$$\mathbf{s}' = \mathbf{s} - \mathbf{1} \bar{s},$$

$$\alpha' = \alpha + \bar{s}$$

Kinship bias affects heritability estimation

Model:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{s} + \epsilon,$$

$$\mathbf{s} + \epsilon \sim \text{Normal}(\mathbf{0}, 2\sigma_G^2\Phi + \sigma_E^2\mathbf{I}).$$

Heritability definition:

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

Variance is estimated with bias:

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

Kinship bias affects heritability estimation

Model:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{s} + \boldsymbol{\epsilon},$$

$$\mathbf{s} + \boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, 2\sigma_G^2\boldsymbol{\Phi} + \sigma_E^2\mathbf{I}).$$

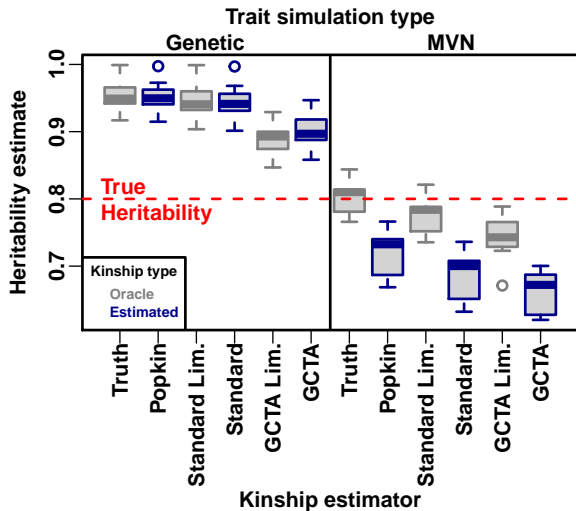
Heritability definition:

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

Variance is estimated with bias:

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

There are more sources of bias!!!



LIGERA (Light GENetic Robust Association): a reversed LMM

Linear mixed-effects model (LMM):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{s} + \boldsymbol{\epsilon}, \quad \mathbf{s} + \boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, 2\sigma_G^2\boldsymbol{\Phi} + \sigma_E^2\mathbf{I}).$$

LIGERA:

$$\mathbf{x}_i = \mathbf{Y}\boldsymbol{\beta} + \mathbf{s}, \quad \mathbf{s} \sim \text{Normal}(\mathbf{0}, \sigma^2\boldsymbol{\Phi}),$$

where here \mathbf{X} , \mathbf{Y} include covariates and intercept.

LIGERA (Light GENetic Robust Association): a reversed LMM

Linear mixed-effects model (LMM):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{s} + \epsilon, \quad \mathbf{s} + \epsilon \sim \text{Normal}(\mathbf{0}, 2\sigma_G^2\Phi + \sigma_E^2I).$$

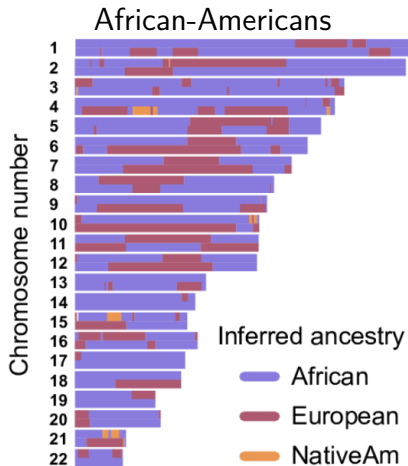
LIGERA:

$$\mathbf{x}_i = \mathbf{Y}\beta + \mathbf{s}, \quad \mathbf{s} \sim \text{Normal}(\mathbf{0}, \sigma^2\Phi),$$

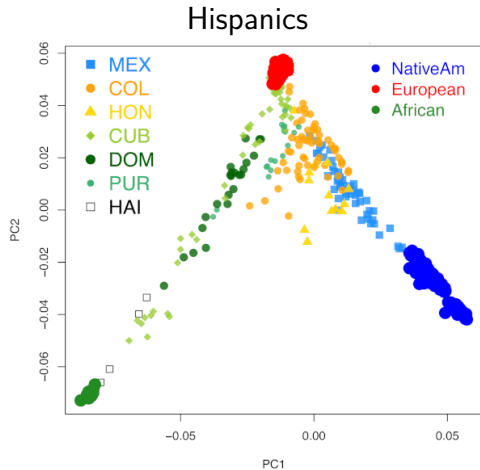
where here \mathbf{X} , \mathbf{Y} include covariates and intercept.

- ▶ LIGERA is faster: no need to fit σ_G^2, σ_E^2 , a slow LMM step!
- ▶ But Standard Estimator is singular, LIGERA requires non-singular Φ

Recently-admixed populations

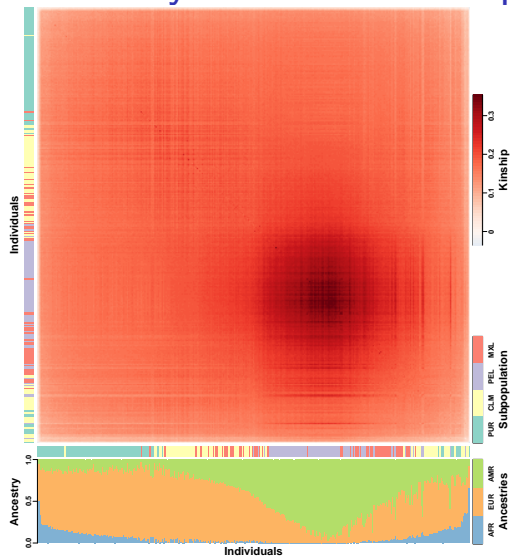


Baharian *et al.* (2016)



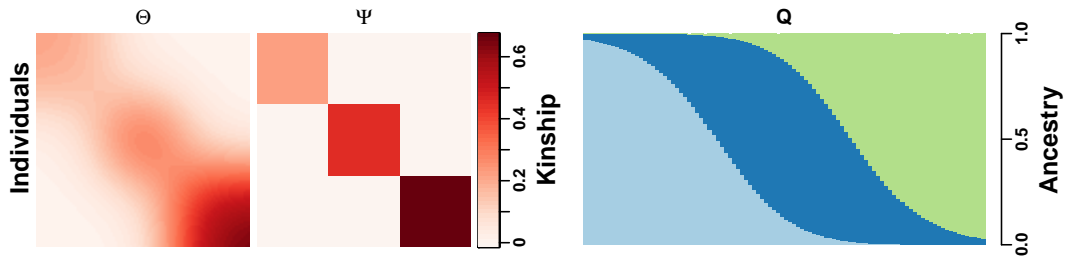
Moreno-Estrada *et al.* (2013)

Population kinship driven by admixture in Hispanics



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

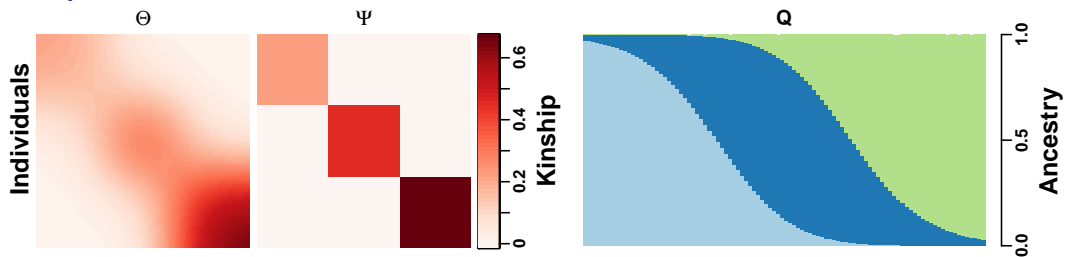
Kinship under the admixture model



$$\Theta = Q\Psi Q^T$$

(Only for unbiased kinship)

Kinship under the admixture model



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

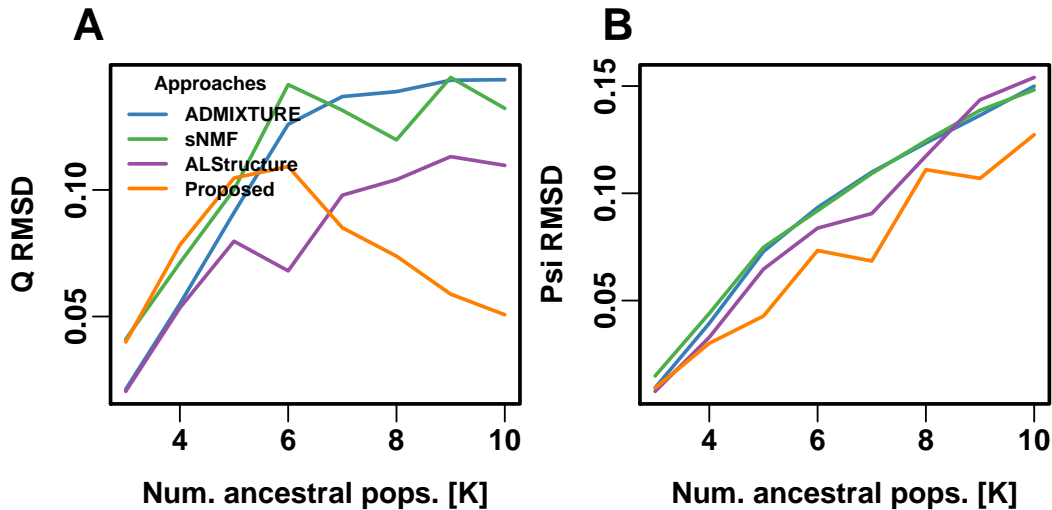
(Only for unbiased kinship)

Can we reverse this formula?

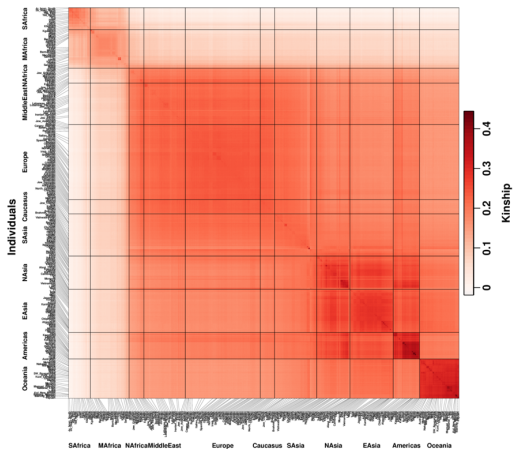
Constrained optimization, regularized objective:

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

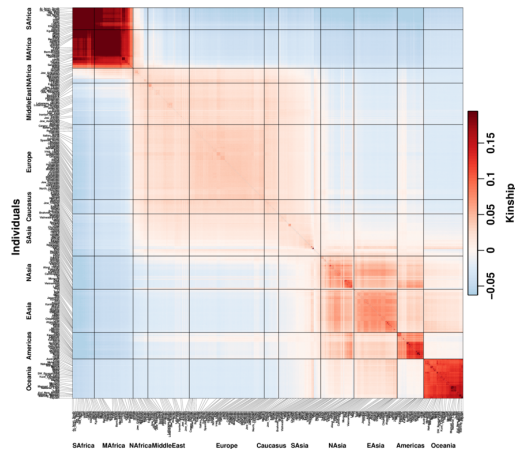
AdmixCor: accuracy



Unbiased kinship estimates: new models, opportunities



New "popkin"
kinship estimator



Biased "standard"
kinship estimator

Acknowledgments

Ochoa Lab

Amika Sood

Tiffany Tu

RP Pornmongkolsuk

Yiqi Yao

Zhuoran Hou

Jiajie Shen

Emmanuel Mokel

Princeton University

John D. Storey

Duke University

Rasheed Gbadegesin

Kouros Owzar

Beth Hauser

Yi-Ju Li

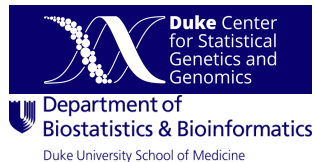
Andrew Allen

Amy Goldberg


Funding

NIH

Whitehead Scholars



 DrAlexOchoa

 ochoalab.github.io

 alejandro.ochoa@duke.edu