Statistical models and methods for human genetic data

Alejandro Ochoa

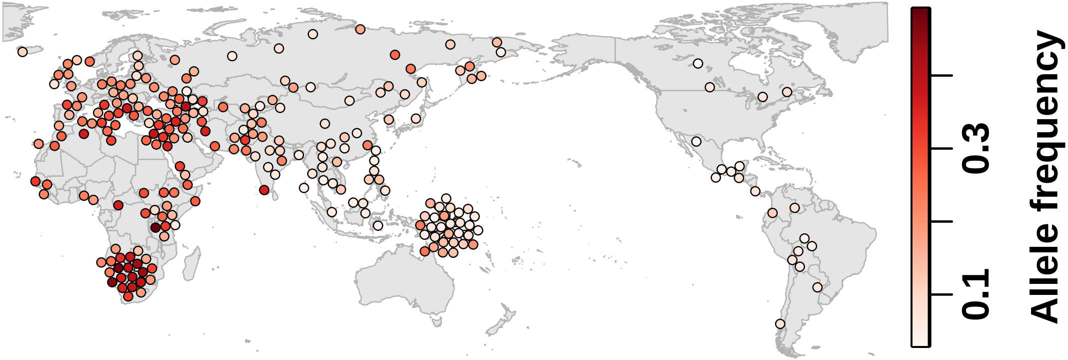
—

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

Biostatistics and Bioinformatics, StatGen — Duke University

2021-03-10 — WSU Mathematics and Statistics

Human genetic structure



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

rs17110306; median differentiation among loci with minor allele frequency $\geq$ 10%

Why? Migration and isolation, admixture, family structure

Overview

New population kinship and F_{ST} estimates

- Human Origins dataset
- Simulation validations

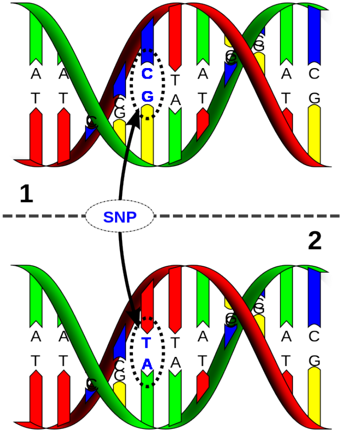
Genetic association models

- Robustness of PCA and LMM approaches
- Biases in heritability estimation
- LIGERA: Light Genetic Robust Association

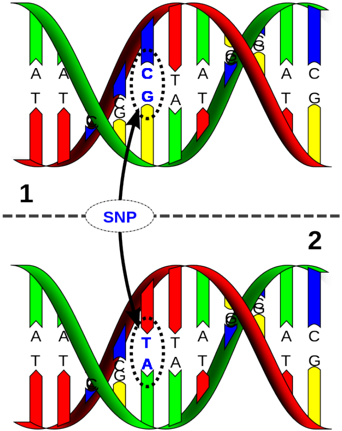
Admixture model

- Hispanics in 1000 Genomes Project
- Joint inference of admixture and population history from genetic covariance

Single Nucleotide Polymorphism (SNP) data



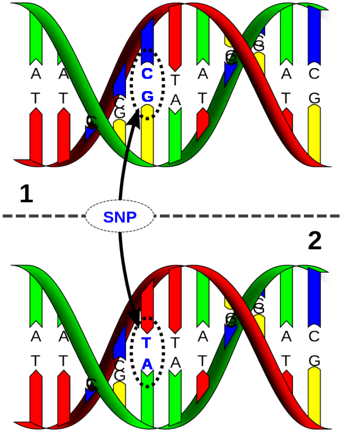
Single Nucleotide Polymorphism (SNP) data



⇒

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 ...

$\mathbf{X}$

Hardy-Weinberg Equillibrium (HWE): Binomial draws

x_{ij} = genotype at locus i for individual j.

p_i = frequency of reference allele at locus i.

Hardy-Weinberg Equillibrium (HWE): Binomial draws

x_{ij} = genotype at locus i for individual j.

p_i = frequency of reference allele at locus i.

Under HWE:

$$\begin{aligned}
\Pr(x_{ij} = 2) &= p_i^2, \\
\Pr(x_{ij} = 1) &= 2p_i\left(1 - p_i\right), \\
\Pr(x_{ij} = 0) &= \left(1 - p_i\right)^2.
\end{aligned}$$

Hardy-Weinberg Equillibrium (HWE): Binomial draws

x_{ij} = genotype at locus i for individual j.

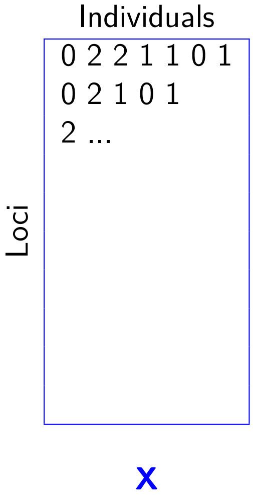
p_i = frequency of reference allele at locus i.

Under HWE:

$$\begin{aligned}\Pr(x_{ij} = 2) &= p_i^2, \\ \Pr(x_{ij} = 1) &= 2p_i\left(1 - p_i\right), \\ \Pr(x_{ij} = 0) &= \left(1 - p_i\right)^2.\end{aligned}$$

HWE not valid under genetic structure!

Goal: measure dependence structure of genotype matrix columns



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

Goal: measure dependence structure of genotype matrix columns

Individuals

Loci

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

X

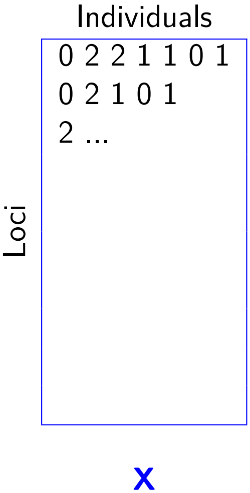
High-dimensional binomial data

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

Relatedness / Population structure

- Dependence between individuals (columns)

Goal: measure dependence structure of genotype matrix columns


Individuals

0 2 2 1 1 0 1

0 2 1 0 1

2 ...

Loci

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)

Model parameters

IBD: “Identical By Descent” (given implicit ancestral pop.) — shared coin flips

Model parameters

IBD: “Identical By Descent” (given implicit ancestral pop.) — shared coin flips

f_j: **Inbreeding coefficient**

Pr. that the two alleles at a random locus of individual j are IBD

$$\mathrm{Var}(x_{ij}) = 2p_i\,(1 - p_i)\,(1 + f_j)$$

Model parameters

IBD: “Identical By Descent” (given implicit ancestral pop.) — shared coin flips

f_j: **Inbreeding coefficient**

Pr. that the two alleles at a random locus of individual j are IBD

$$\mathrm{Var}(x_{ij}) = 2p_i\,(1-p_i)\,(1+f_j)$$

φ_{jk}: **Kinship coefficient**

Pr. that two alleles, one at random from each of individuals j and k, at one random locus are IBD

$$\mathrm{Cov}(x_{ij}, x_{ik}) = 4p_i\,(1-p_i)\,\varphi_{jk}$$

Model parameters

IBD: “Identical By Descent” (given implicit ancestral pop.) — shared coin flips

f_j: **Inbreeding coefficient**

Pr. that the two alleles at a random locus of individual j are IBD

$$\mathrm{Var}(x_{ij}) = 2p_i(1-p_i)(1+f_j)$$

φ_{jk}: **Kinship coefficient**

Pr. that two alleles, one at random from each of individuals j and k, at one random locus are IBD

$$\mathrm{Cov}(x_{ij}, x_{ik}) = 4p_i(1-p_i)\varphi_{jk}$$

F_{ST}: **Fixation index**

Pr. that two random alleles in a **subpopulation** at a random locus are IBD

Overview

New population kinship and F_{ST} estimates

- **Human Origins dataset**
- **Simulation validations**

Genetic association models

- Robustness of PCA and LMM approaches
- Biases in heritability estimation
- LIGERA: Light Genetic Robust Association

Admixture model

- Hispanics in 1000 Genomes Project
- Joint inference of admixture and population history from genetic covariance

New kinship estimator for general relatedness

New kinship estimator for general relatedness

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

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

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**:

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

New kinship 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**:

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

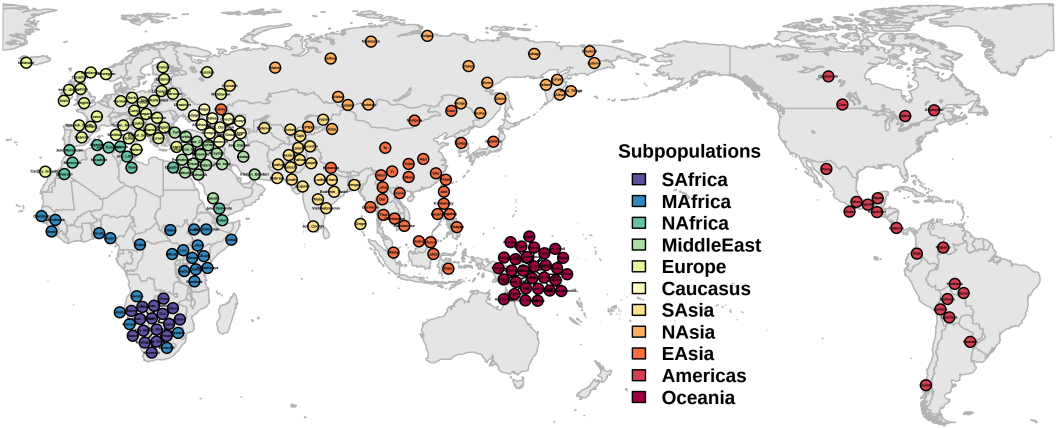
`popkin`: first unbiased kinship estimator! R package (Ochoa and Storey, 2021)

$$A_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij} - 1)(x_{ik} - 1) - 1, \qquad \hat{A}_{\min} = \min_{u \neq v} \frac{1}{|S_u||S_v|}\sum_{j \in S_u}\sum_{k \in S_v} A_{jk},$$

$$\hat{\varphi}_{jk}^{\mathrm{new}} = 1 - \frac{A_{jk}}{\hat{A}_{\min}} \xrightarrow[m \to \infty]{\text{a.s.}} \varphi_{jk}.$$

https://github.com/StoreyLab/popkin

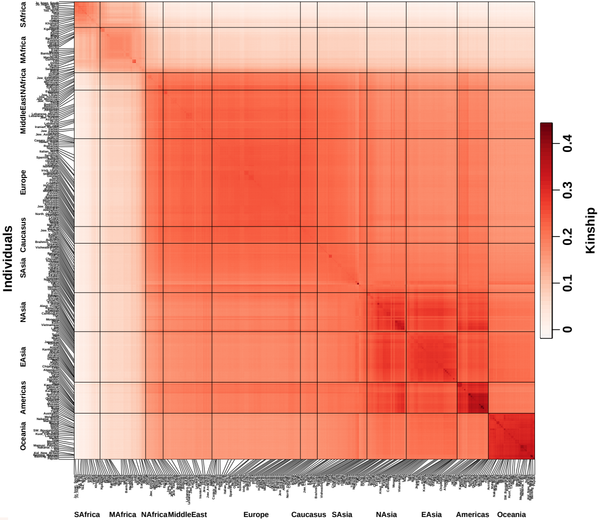
Dataset: Human Origins

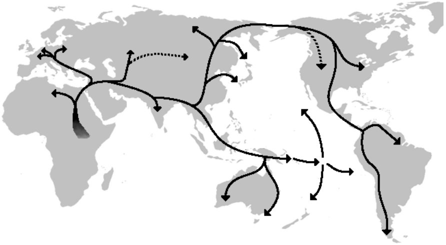


Lazaridis *et al.* (2014), (2016); Skoglund *et al.* (2016)

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

Kinship matrix of world-wide human population

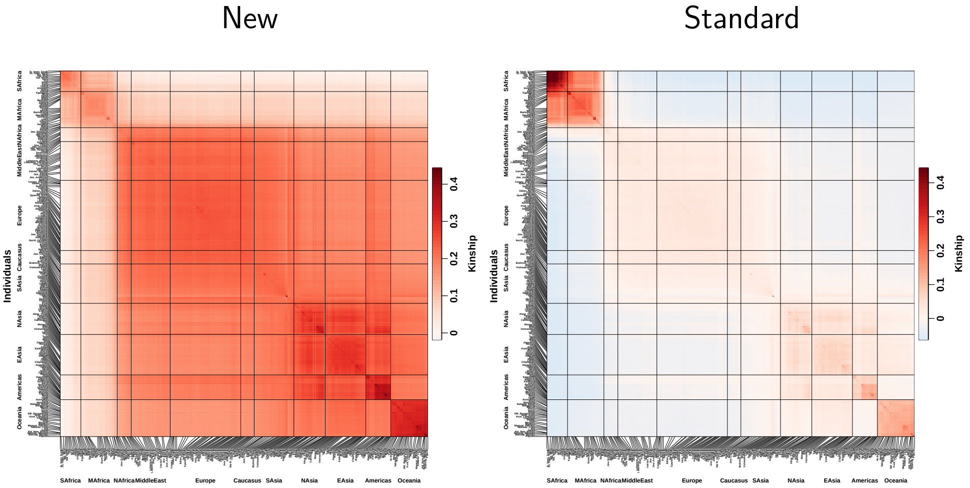




Ochoa and Storey (2019) doi:10.1101/653279

https://github.com/StoreyLab/popkin

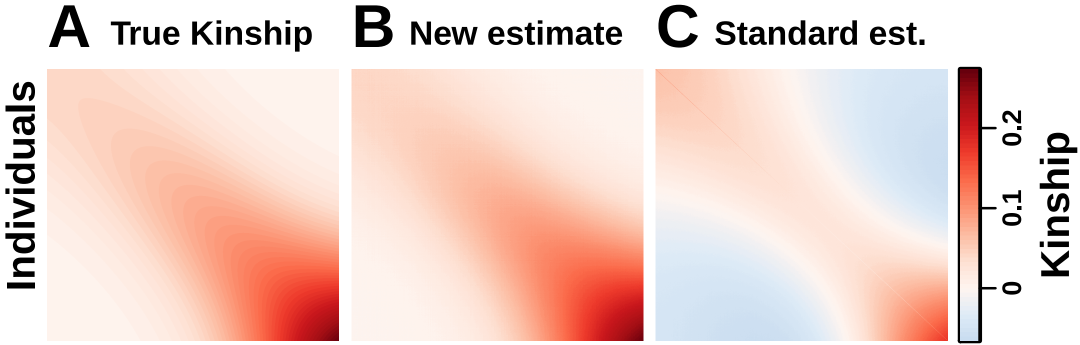
Standard kinship estimator is severely biased



Ochoa and Storey (2019) doi:10.1101/653279

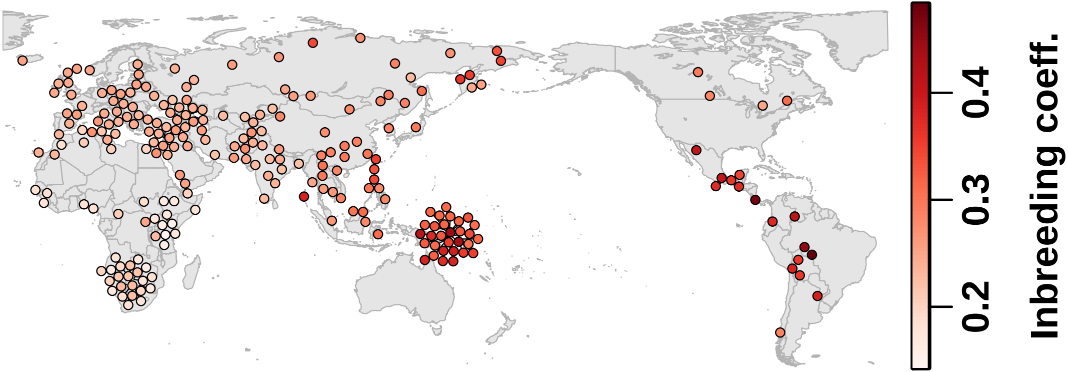
https://github.com/StoreyLab/popkin

Validation in simulation



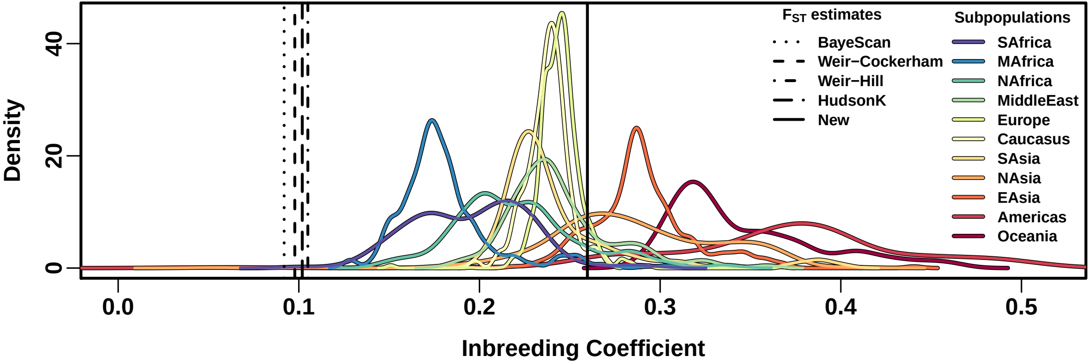
Ochoa and Storey (2021) doi:10.1371/journal.pgen.1009241

Population-level inbreeding increases with distance from Africa



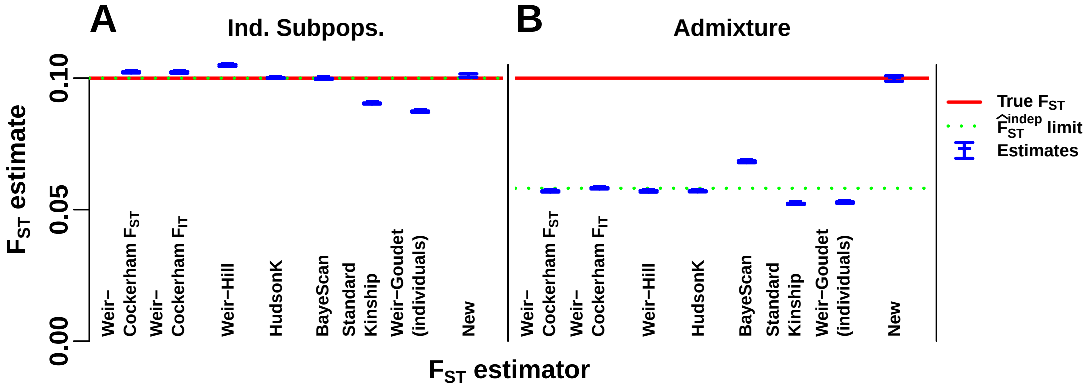
Ochoa and Storey (2019) doi:10.1101/653279

Differentiation (F_{ST}) previously underestimated



Ochoa and Storey (2019) doi:10.1101/653279

Validation in simulation



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

Overview

New population kinship and F_{ST} estimates

- Human Origins dataset
- Simulation validations

Genetic association models

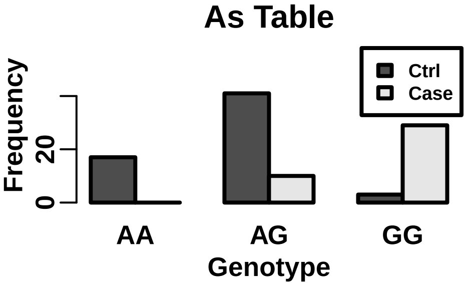
- **Robustness of PCA and LMM approaches**
- **Biases in heritability estimation**
- **LIGERA: Light Genetic Robust Association**

Admixture model

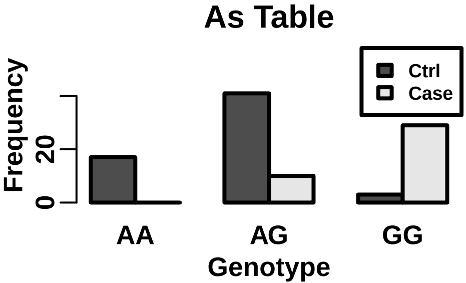
- Hispanics in 1000 Genomes Project
- Joint inference of admixture and population history from genetic covariance

Genetic association study: genotype-phenotype correlation

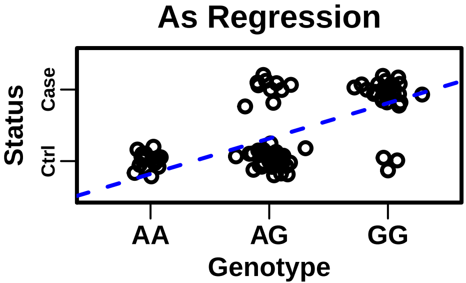
Genetic association study: genotype-phenotype correlation



Genetic association study: genotype-phenotype correlation



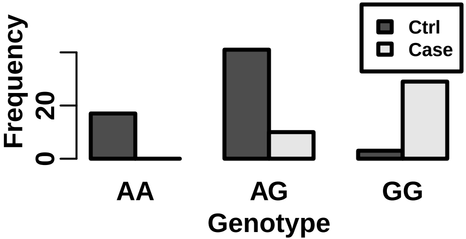
As Table



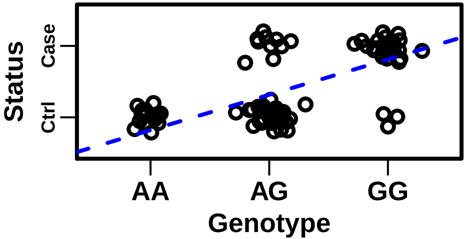
As Regression

Genetic association study: genotype-phenotype correlation

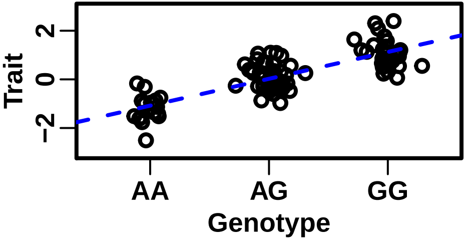
As Table



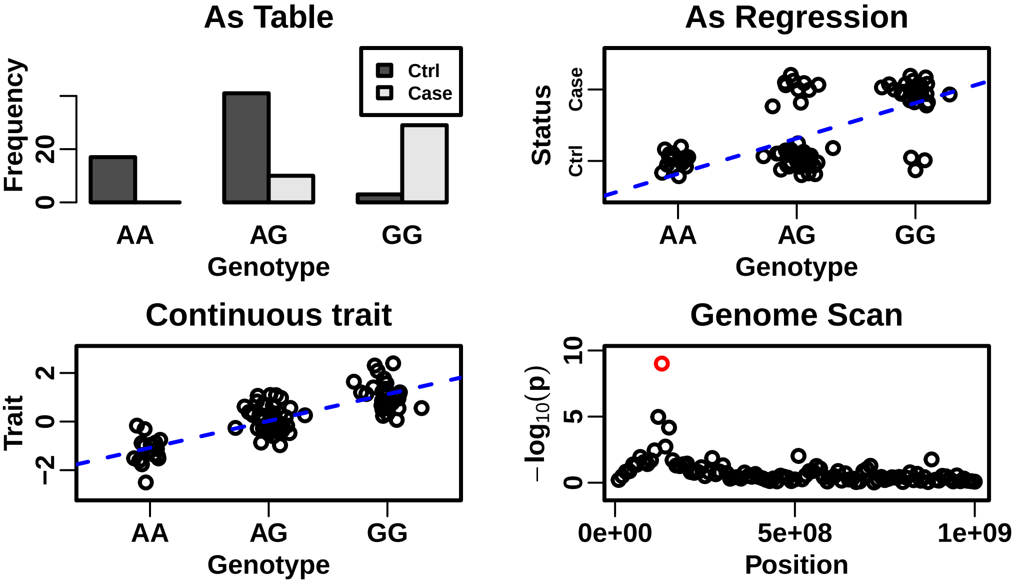
As Regression



Continuous trait



Genetic association study: genotype-phenotype correlation



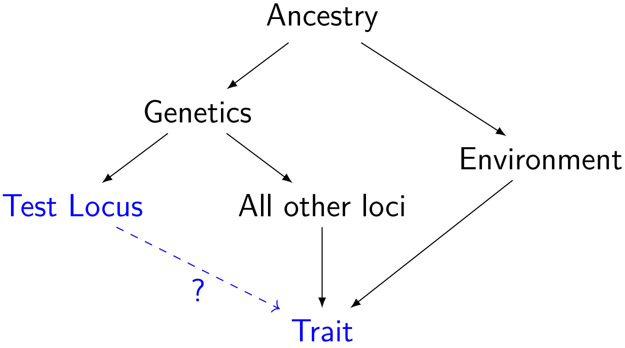
Why is this problem so hard?

Why is this problem so hard?

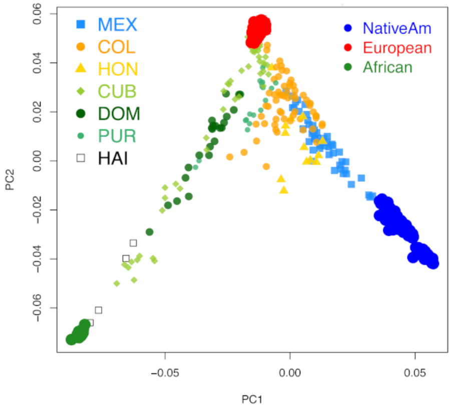
- Millions of tests
- Polygenicity
- Confounders

Why is this problem so hard?

- Millions of tests
- Polygenicity
- Confounders

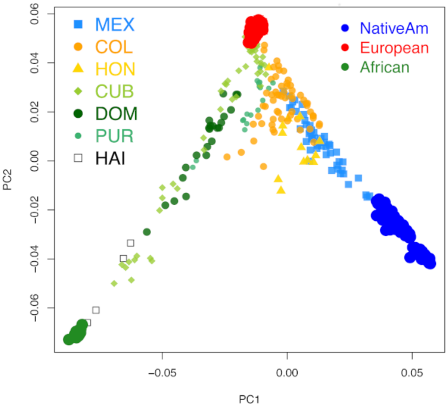


PCA: Principal Component Analysis



Moreno-Estrada *et al.* (2013)

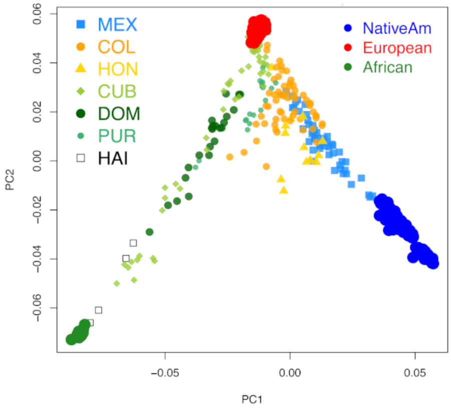
PCA: Principal Component Analysis



Moreno-Estrada *et al.* (2013)

PCs map to ancestry.

PCA: Principal Component Analysis

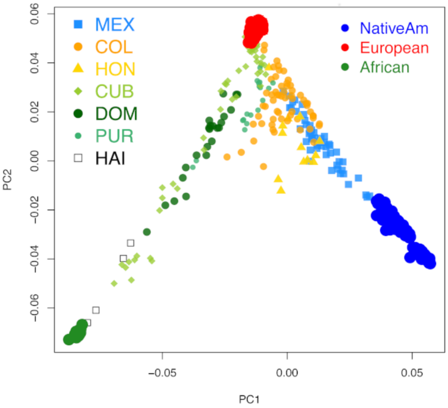


Moreno-Estrada *et al.* (2013)

PCs map to ancestry.

"PCs" are top eigenvectors of kinship matrix.

PCA: Principal Component Analysis



Moreno-Estrada *et al.* (2013)

PCs map to ancestry.

"PCs" are top eigenvectors of kinship matrix.

Pros: Fast!

Cons: Fails on family data.

Genetic association methods: PCA and LMM

Principal components analysis (PCA) association model: fixed-effects regression:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{U}_d\gamma_d + \epsilon.$$

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

Genetic association methods: PCA and LMM

Principal components analysis (PCA) association model: fixed-effects regression:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{U}_d\gamma_d + \epsilon.$$

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

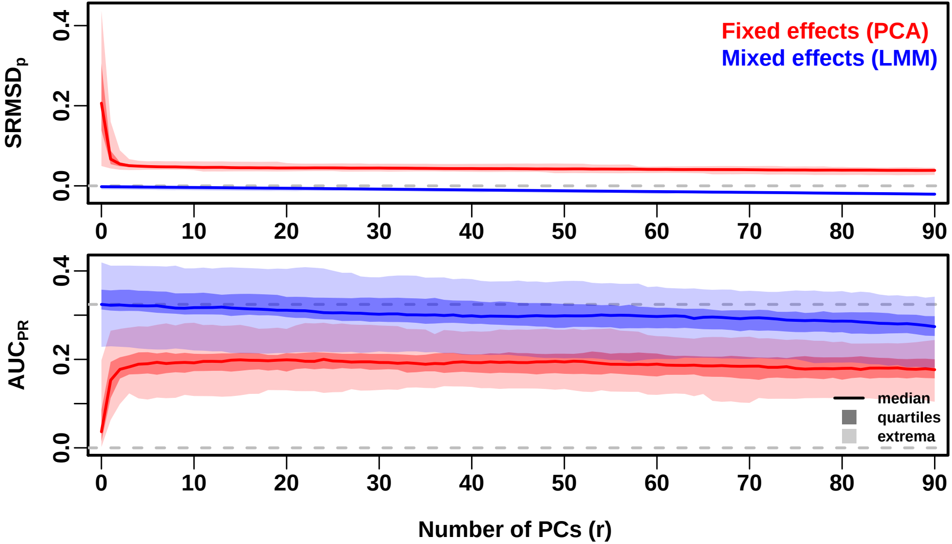
Linear mixed-effects model (LMM):

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

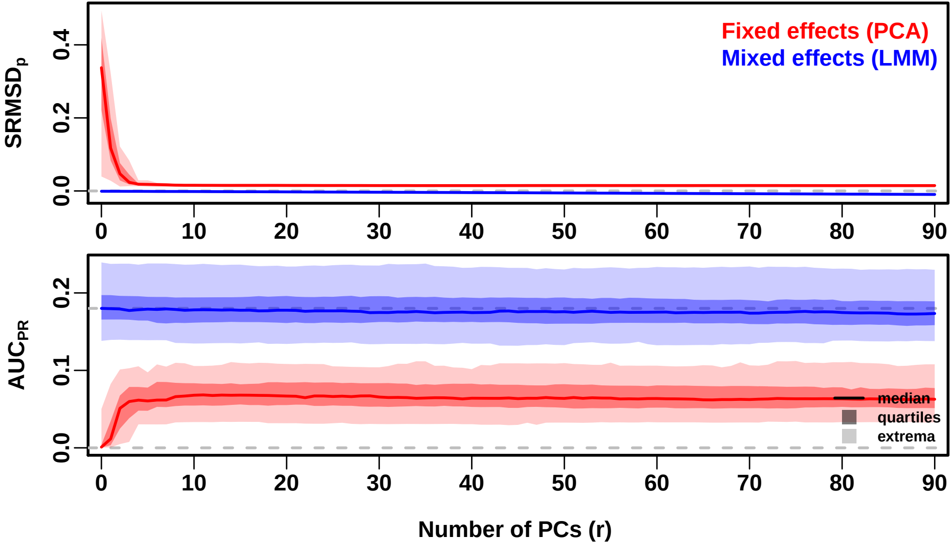
- Random effect has covariance structure from kinship matrix $\mathbf{\Phi}$:

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

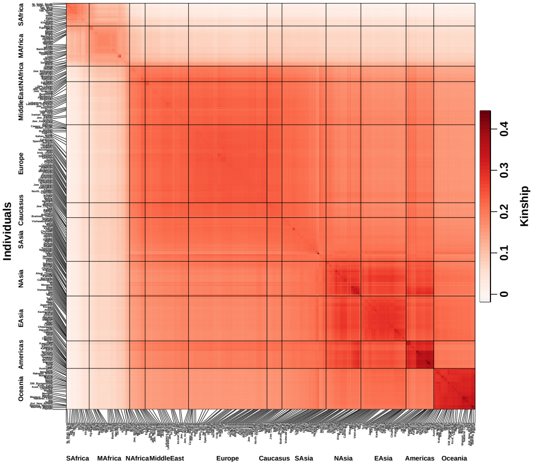
LMM outperforms PCA: Simulated admixture + family structure



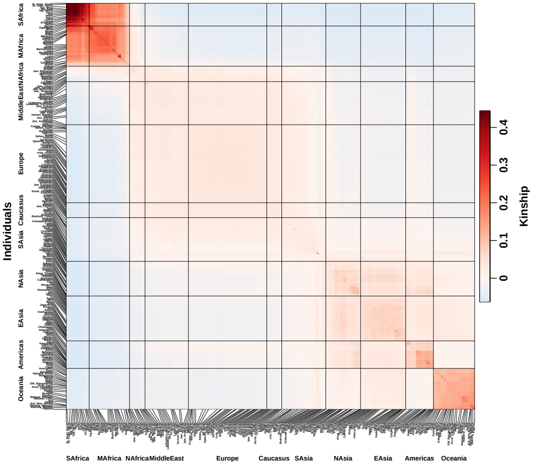
LMM outperforms PCA: 1000 Genomes Project + sim trait



Kinship bias does not affect genetic associations



New popkin
kinship estimator



Standard
kinship estimator

Kinship bias does not affect genetic associations

Centering matrix is key to understanding kinship bias algebraically:

$$\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{J}.$$

Kinship bias does not affect genetic associations

Centering matrix is key to understanding kinship bias algebraically:

$$\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{J}.$$

Standard kinship bias as a transformation of true kinship by centering:

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

Kinship bias does not affect genetic associations

Centering matrix is key to understanding kinship bias algebraically:

$$\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{J}.$$

Standard kinship bias as a transformation of true kinship by centering:

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

Matrix square root also centered:

$$(\mathbf{\Phi}')^{\frac{1}{2}} = \frac{1}{\sqrt{1 - \bar{\varphi}}}\mathbf{C}\mathbf{\Phi}^{\frac{1}{2}}.$$

Kinship bias does not affect genetic associations

LMM equivalent models:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{s} + \epsilon, \qquad \mathbf{s} \sim \text{Normal}\left(\mathbf{0}, \sigma^2\mathbf{\Phi}\right),$$

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \sigma\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r} + \epsilon, \qquad \mathbf{r} \sim \text{Normal}\left(\mathbf{0}, \mathbf{I}\right).$$

Kinship bias does not affect genetic associations

LMM equivalent models:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{s} + \epsilon, \qquad \mathbf{s} \sim \text{Normal}\left(\mathbf{0}, \sigma^2 \mathbf{\Phi}\right),$$

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \sigma \mathbf{\Phi}^{\frac{1}{2}} \mathbf{r} + \epsilon, \qquad \mathbf{r} \sim \text{Normal}\left(\mathbf{0}, \mathbf{I}\right).$$

Fit under true kinship ($\mathbf{\Phi}$) vs biased limit ($\mathbf{\Phi}'$) is equally good (algebra depends on centering matrix properties):

$$\begin{aligned} \mathbf{y} &= \mathbf{1}\alpha + \mathbf{x}_i\beta + \sigma \mathbf{\Phi}^{\frac{1}{2}} \mathbf{r} + \epsilon \\ &= \mathbf{1}\alpha' + \mathbf{x}_i\beta' + \sigma' \left(\mathbf{\Phi}'\right)^{\frac{1}{2}} \mathbf{r}' + \epsilon', \end{aligned}$$

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

Kinship bias does not affect genetic associations

LMM equivalent models:

$$\begin{aligned} \mathbf{y} &= \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{s} + \epsilon, & \mathbf{s} &\sim \text{Normal}\left(\mathbf{0}, \sigma^2\mathbf{\Phi}\right), \\ \mathbf{y} &= \mathbf{1}\alpha + \mathbf{x}_i\beta + \sigma\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r} + \epsilon, & \mathbf{r} &\sim \text{Normal}\left(\mathbf{0}, \mathbf{I}\right). \end{aligned}$$

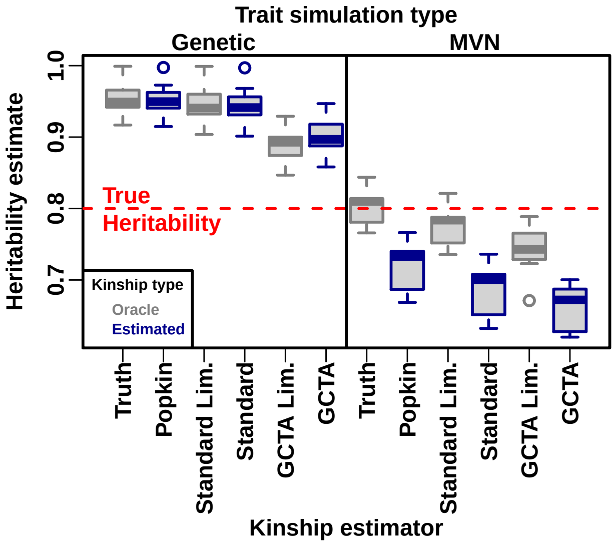
Fit under true kinship $(\mathbf{\Phi})$ vs biased limit $(\mathbf{\Phi}')$ is equally good (algebra depends on centering matrix properties):

$$\begin{aligned} \mathbf{y} &= \mathbf{1}\alpha + \mathbf{x}_i\beta + \sigma\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r} + \epsilon \\ &= \mathbf{1}\alpha' + \mathbf{x}_i\beta' + \sigma'\left(\mathbf{\Phi}'\right)^{\frac{1}{2}}\mathbf{r}' + \epsilon', \end{aligned}$$

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

Similar argument holds approximately for PCA regression.

Kinship bias affects heritability estimation



Future work: Tuning elastic nets for genetic association

$$\hat{\beta} \equiv \underset{\beta}{\mathbf{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

- Validate existing PCA extension
- How to model higher-dimensional relatedness?

Genetic association models: forward vs reversed

(Forward) linear mixed-effects model	Reverse model

Genetic association models: forward vs reversed

(Forward) linear mixed-effects model	Reverse model
$\mathbf{y} = \alpha_i + \mathbf{x}_i \beta_i + \mathbf{F}\gamma_i + \mathbf{r}_i$	

Genetic association models: forward vs reversed

(Forward) linear mixed-effects model	Reverse model
$\mathbf{y} = \alpha_i + \mathbf{x}_i\beta_i + \mathbf{F}\gamma_i + \mathbf{r}_i$	$\mathbf{x}_i = \alpha'_i + \mathbf{y}\beta'_i + \epsilon'_i$

Genetic association models: forward vs reversed

(Forward) linear mixed-effects model	Reverse model
$\mathbf{y} = \alpha_i + \mathbf{x}_i\beta_i + \mathbf{F}\gamma_i + \mathbf{r}_i$	$\mathbf{x}_i = \alpha'_i + \mathbf{y}\beta'_i + \epsilon'_i$
Models trait (complicated, unknowns)	

Genetic association models: forward vs reversed

(Forward) linear mixed-effects model	Reverse model
$\mathbf{y} = \alpha_i + \mathbf{x}_i \beta_i + \mathbf{F}\gamma_i + \mathbf{r}_i$	$\mathbf{x}_i = \alpha'_i + \mathbf{y}\beta'_i + \epsilon'_i$
Models trait (complicated, unknowns)	Models genotype (kinship).

Genetic association models: forward vs reversed

(Forward) linear mixed-effects model	Reverse model
$\mathbf{y} = \alpha_i + \mathbf{x}_i \beta_i + \mathbf{F} \gamma_i + \mathbf{r}_i$	$\mathbf{x}_i = \alpha'_i + \mathbf{y} \beta'_i + \epsilon'_i$
Models trait (complicated, unknowns)	Models genotype (kinship).
	Environment can be absent (Song, Hao, Storey 2015)

Genetic association models: forward vs reversed

(Forward) linear mixed-effects model	Reverse model
$\mathbf{y} = \alpha_i + \mathbf{x}_i \beta_i + \mathbf{F} \gamma_i + \mathbf{r}_i$	$\mathbf{x}_i = \alpha'_i + \mathbf{y} \beta'_i + \epsilon'_i$
Models trait (complicated, unknowns)	Models genotype (kinship).
	Environment can be absent (Song, Hao, Storey 2015)
Random effects are slow!	

Genetic association models: forward vs reversed

(Forward) linear mixed-effects model	Reverse model
$\mathbf{y} = \alpha_i + \mathbf{x}_i \beta_i + \mathbf{F}\gamma_i + \mathbf{r}_i$	$\mathbf{x}_i = \alpha'_i + \mathbf{y}\beta'_i + \epsilon'_i$
Models trait (complicated, unknowns)	Models genotype (kinship).
	Environment can be absent (Song, Hao, Storey 2015)
Random effects are slow!	Fast!

LIGERA: light genetic robust association

Objective function: move genetic structure to residuals:

$$G = \left(\mathbf{Y} \beta_i - \mathbf{x}_i \right)^\intercal \mathbf{\Phi}^{-1} \left(\mathbf{Y} \beta_i - \mathbf{x}_i \right).$$

LIGERA: light genetic robust association

Objective function: move genetic structure to residuals:

$$G = \left(\mathbf{Y}\beta_i - \mathbf{x}_i\right)^\intercal \mathbf{\Phi}^{-1} \left(\mathbf{Y}\beta_i - \mathbf{x}_i\right).$$

Effect size estimator is matrix product of data:

$$\hat{\beta}_i = \mathbf{H}^\intercal \mathbf{x}_i, \qquad \mathbf{H} = \mathbf{\Phi}^{-1}\mathbf{Y}\left(\mathbf{Y}^\intercal \mathbf{\Phi}^{-1}\mathbf{Y}\right)^{-1}.$$

LIGERA: light genetic robust association

Objective function: move genetic structure to residuals:

$$G = \left(\mathbf{Y}\beta_i - \mathbf{x}_i\right)^\intercal \mathbf{\Phi}^{-1} \left(\mathbf{Y}\beta_i - \mathbf{x}_i\right).$$

Effect size estimator is matrix product of data:

$$\hat{\beta}_i = \mathbf{H}^\intercal \mathbf{x}_i, \qquad \mathbf{H} = \mathbf{\Phi}^{-1}\mathbf{Y}\left(\mathbf{Y}^\intercal \mathbf{\Phi}^{-1}\mathbf{Y}\right)^{-1}.$$

Variance under null hypothesis has closed form:

$$\mathrm{Cov}\left(\hat{\beta}_i \middle| \mathbf{Y}\right) = 4p_i\left(1 - p_i\right)\left(\mathbf{H}^\intercal \mathbf{\Phi} \mathbf{H}\right), \qquad \left(\mathbf{H}^\intercal \mathbf{\Phi} \mathbf{H}\right) = \left(\mathbf{Y}^\intercal \mathbf{\Phi}^{-1}\mathbf{Y}\right)^{-1}.$$

LIGERA: light genetic robust association

Objective function: move genetic structure to residuals:

$$G = (\mathbf{Y}\beta_i - \mathbf{x}_i)^\intercal \mathbf{\Phi}^{-1} (\mathbf{Y}\beta_i - \mathbf{x}_i) \, .$$

Effect size estimator is matrix product of data:

$$\hat{\beta}_i = \mathbf{H}^\intercal \mathbf{x}_i, \qquad \mathbf{H} = \mathbf{\Phi}^{-1}\mathbf{Y}\left(\mathbf{Y}^\intercal \mathbf{\Phi}^{-1}\mathbf{Y}\right)^{-1} .$$

Variance under null hypothesis has closed form:

$$\mathrm{Cov}\left(\hat{\beta}_i \middle| \mathbf{Y}\right) = 4p_i(1-p_i)(\mathbf{H}^\intercal \mathbf{\Phi} \mathbf{H}) \, , \qquad (\mathbf{H}^\intercal \mathbf{\Phi} \mathbf{H}) = \left(\mathbf{Y}^\intercal \mathbf{\Phi}^{-1}\mathbf{Y}\right)^{-1} .$$

This is fast! Bottleneck is calculating $\mathbf{\Phi}^{-1}\mathbf{Y}$.
Solve efficiently with “conjugate gradient” algorithm!

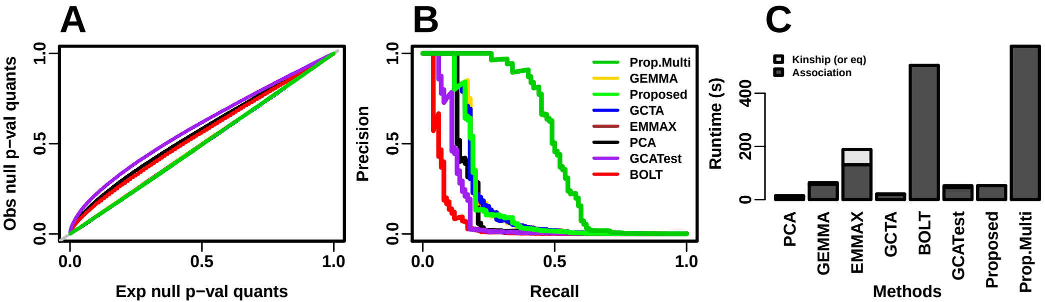
LIGERA: light genetic robust association

“Multiscan”: forward variable selection.

At each iteration:

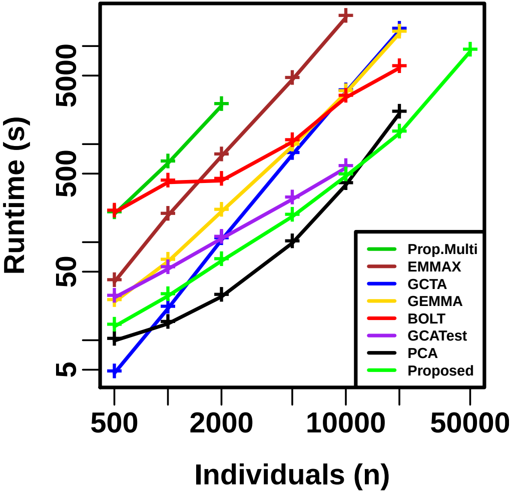
- Calculate p-values
- Estimate q-values
- Select all loci with $q < 0.05$, add as covariates in following iteration
- Stop if no new loci are selected.

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



Overview

New population kinship and F_{ST} estimates

- Human Origins dataset
- Simulation validations

Genetic association models

- Robustness of PCA and LMM approaches
- Biases in heritability estimation
- LIGERA: Light Genetic Robust Association

Admixture model

- Hispanics in 1000 Genomes Project
- Joint inference of admixture and population history from genetic covariance

Acknowledgments

Ochoa Lab

Amika Sood
Zhuoran Hou
Jiajie Shen
Yiqi Yao (now at IQVIA Beijing)

Duke University

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

Princeton University

John D. Storey

Funding

NIH
Duke 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