Modeling Relatedness in Genetic Association Studies

Alejandro Ochoa

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

StatGen, CBB, B&B — Duke University

2019-09-28 — CBB Retreat

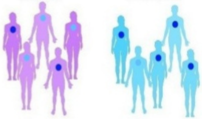
Why study relatedness?

Why study relatedness?

Human genetics is fascinating!

Why study relatedness?

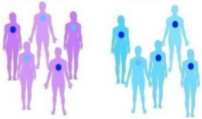
Human genetics is fascinating!



Genetic Association Studies confounded by relatedness

Why study relatedness?

Human genetics is fascinating!

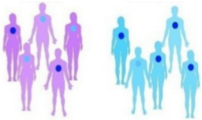


Genetic Association Studies confounded by relatedness

Heritability of complex traits

Why study relatedness?

Human genetics is fascinating!



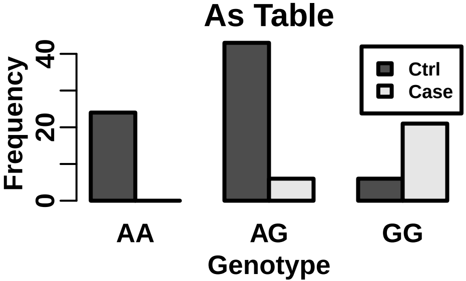
Genetic Association Studies confounded by relatedness

Heritability of complex traits

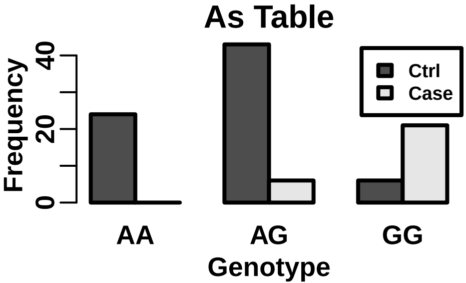
Selection scans

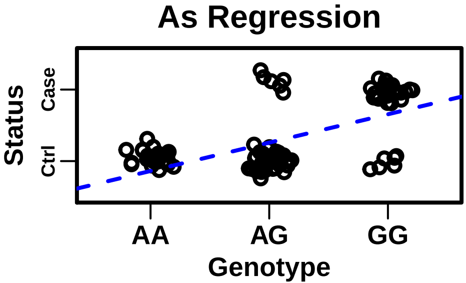
Genetic association study: genotype-phenotype correlation

Genetic association study: genotype-phenotype correlation

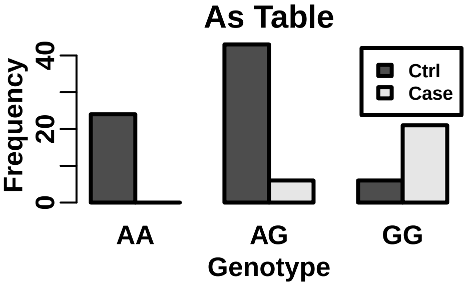


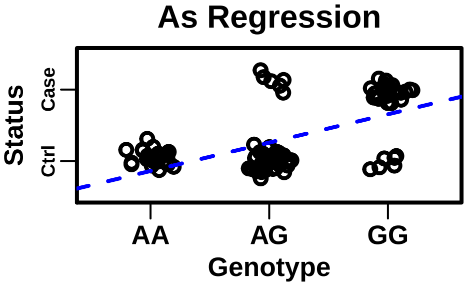
Genetic association study: genotype-phenotype correlation

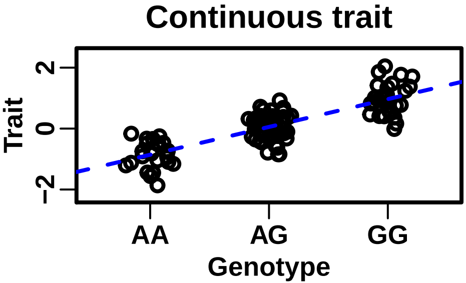




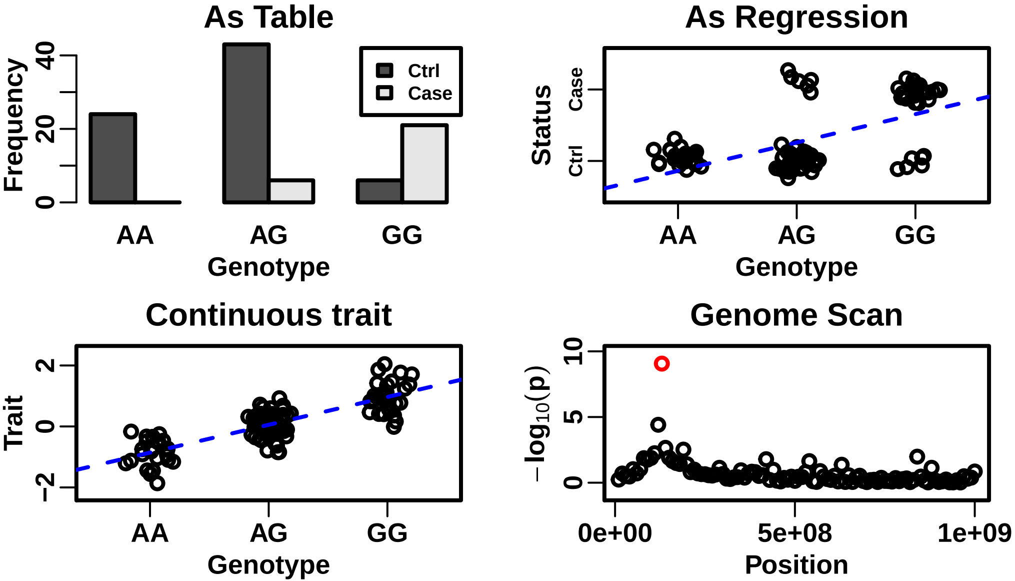
Genetic association study: genotype-phenotype correlation







Genetic association study: genotype-phenotype correlation



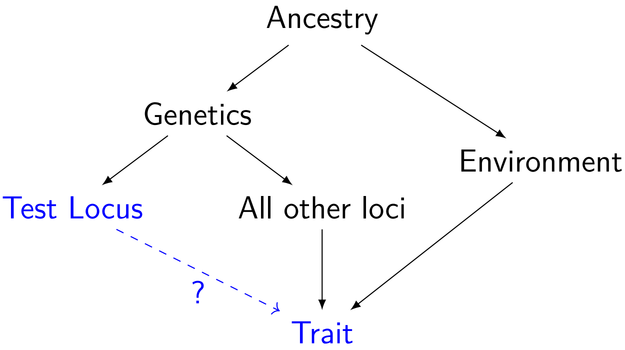
Why is this problem so hard?

Why is this problem so hard?

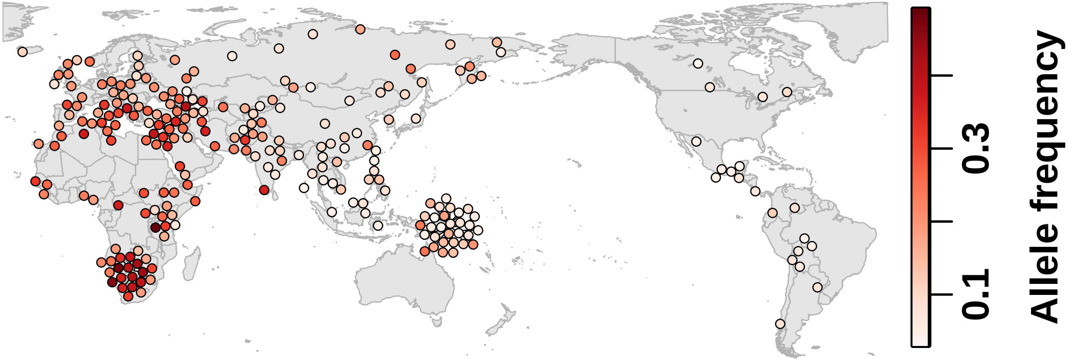
- Millions of tests
- Polygenicity
- Confounders

Why is this problem so hard?

- Millions of tests
- Polygenicity
- Confounders



Median-differentiation human locus



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

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

Basic association tests assume equal allele frequency within cases and controls!

Overview of results

Part I:

- Estimating relatedness under both ancestry and family structure

Part II:

- New approach to genetic association: fast and accurate

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 $x_{ij} \in \{0, 1, 2\}$:

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

Standard estimator is **biased**:

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

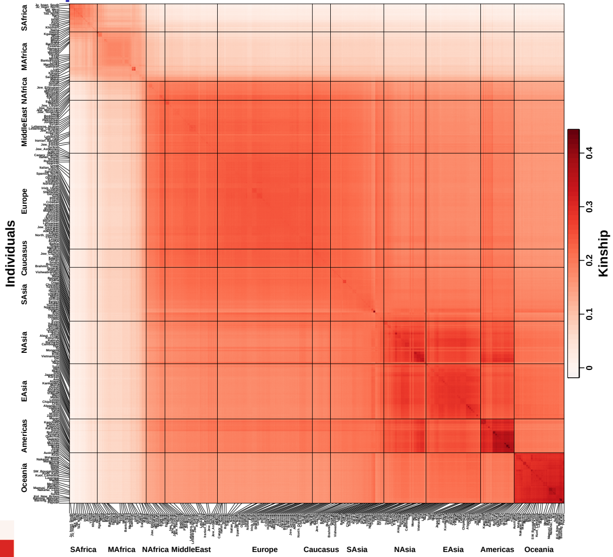
`popkin`: first unbiased kinship estimator! — R package on CRAN

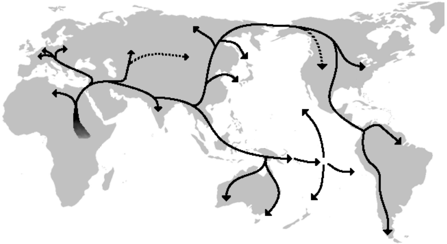
$$A_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij} - 1)(x_{ik} - 1) - 1, \qquad \hat{A}_{\text{min}} = \min_{u \neq v}\frac{1}{|S_u||S_v|}\sum_{j \in S_u}\sum_{k \in S_v} A_{jk},$$

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

https://github.com/StoreyLab/popkin

Kinship matrix of world-wide human population



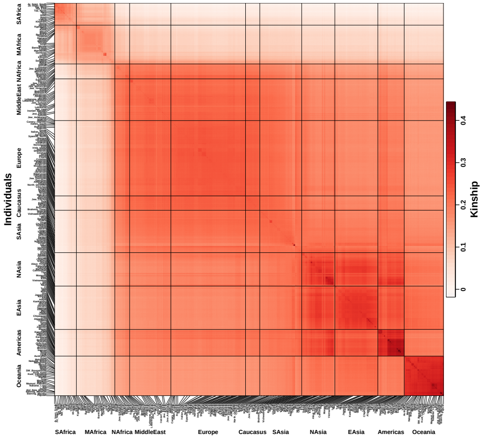


Ochoa and Storey (2019b) doi:10.1101/653279

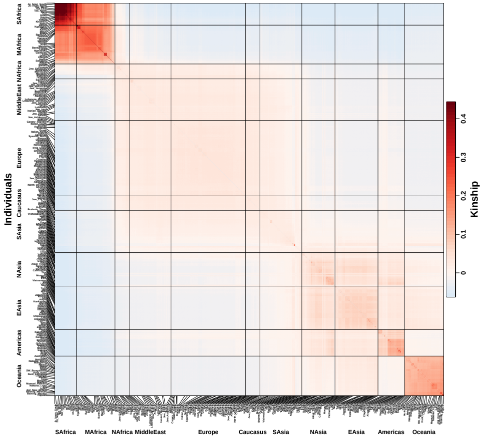
https://github.com/StoreyLab/popkin

Standard kinship estimator is severely biased

New



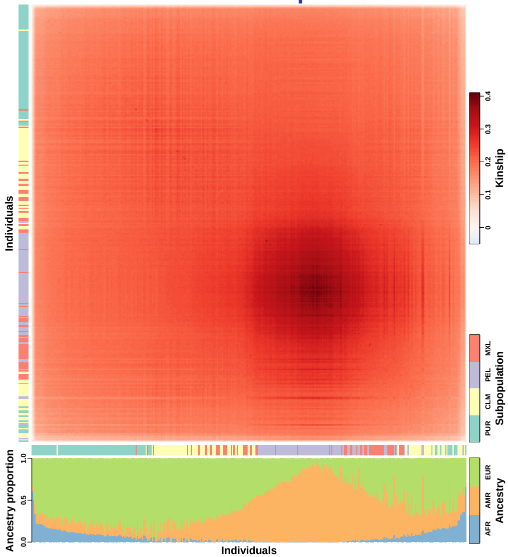
Standard



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

https://github.com/StoreyLab/popkin

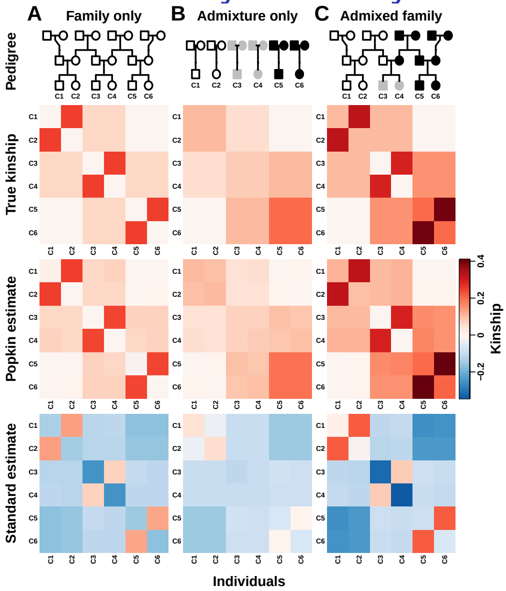
Kinship driven by admixture in Hispanics



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

https://github.com/StoreyLab/popkin

Unified kinship model: ancestry + family structure!



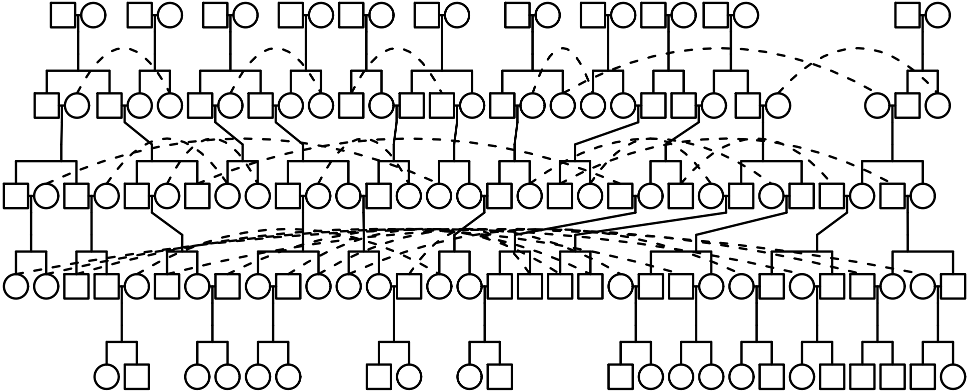
Large family simulation

Mating: proximity preference, but $\leq$ 2nd-cousin avoidance.

Large family simulation

Mating: proximity preference, but ≤ 2nd-cousin avoidance.

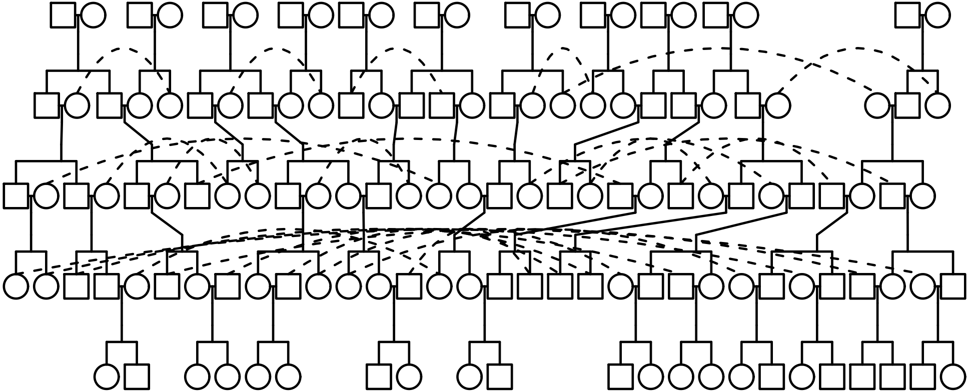
Small example: 22 individuals, 5 generations:



Large family simulation

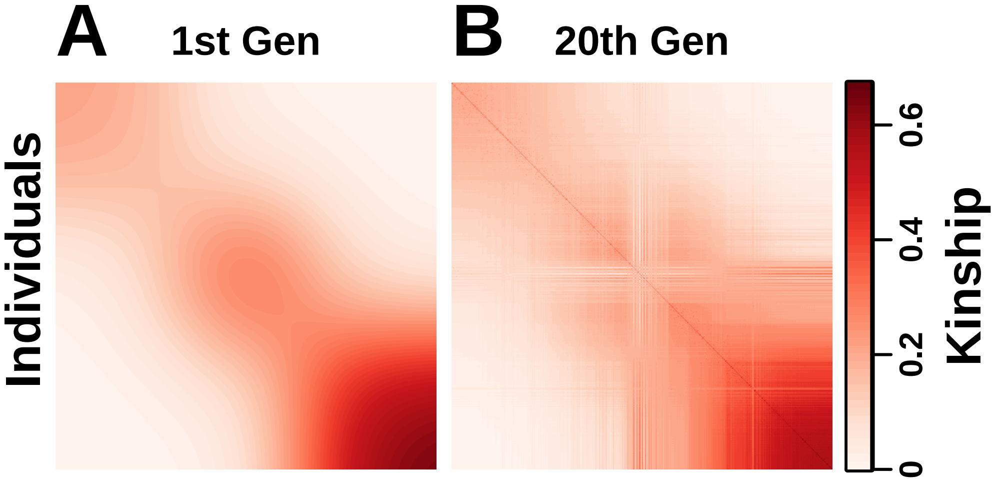
Mating: proximity preference, but ≤ 2nd-cousin avoidance.

Small example: 22 individuals, 5 generations:



Actual simulation: 1000 individuals, 20 generations!

Large family simulation



BNPSD https://github.com/StoreyLab/bnpsd

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)
Authors frequently separate ancestry (PCs) from family (random effect)	

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)
Authors frequently separate ancestry (PCs) from family (random effect)	Unified kinship model

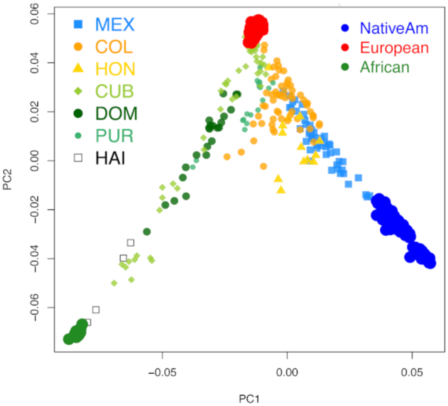
Genetic association models: forward vs reversed

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

Genetic association models: forward vs reversed

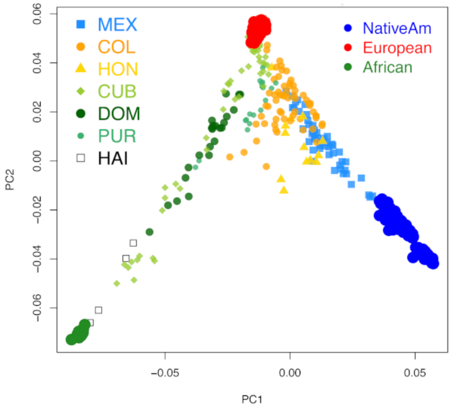
(Forward) linear mixed-effects model	Reverse model
$\mathbf{y} = \alpha_i + \mathbf{x}_i \beta_i + \mathbf{F}\gamma_i + \mathbf{r}_i$	$\mathbf{x}_i = \alpha'_i + \mathbf{y}\beta'_i + \epsilon'_i$
Models trait (complicated, unknowns)	Models genotype (kinship).
	Environment can be absent (Song, Hao, Storey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

PCA: Principal Component Analysis



Moreno-Estrada *et al.* (2013)

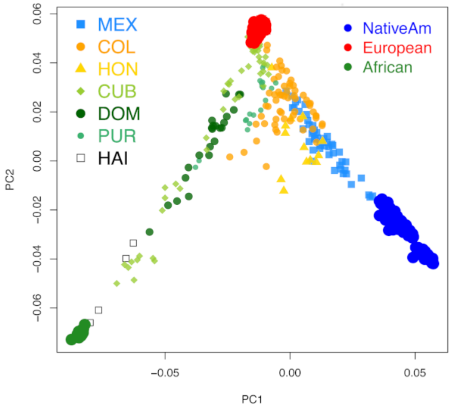
PCA: Principal Component Analysis



Moreno-Estrada *et al.* (2013)

PCs map to ancestry.

PCA: Principal Component Analysis

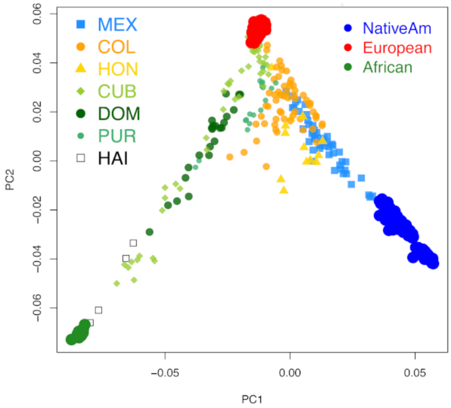


Moreno-Estrada *et al.* (2013)

PCs map to ancestry.

"PCs" are top eigenvectors of kinship matrix.

PCA: Principal Component Analysis



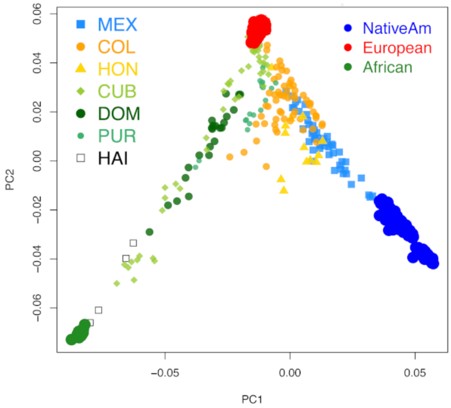
Moreno-Estrada *et al.* (2013)

PCs map to ancestry.

"PCs" are top eigenvectors of kinship matrix.

Pros: Fast! Robust! Accurate when close relatives are excluded.

PCA: Principal Component Analysis



Moreno-Estrada *et al.* (2013)

PCs map to ancestry.

"PCs" are top eigenvectors of kinship matrix.

Pros: Fast! Robust! Accurate when close relatives are excluded.

Cons: Fails on family data.

LMM: Linear Mixed-effects Model

LMM: Linear Mixed-effects Model

“Mixed”: Fixed and Random effects:

$$\begin{aligned}\mathbf{y} &= \alpha_i + \mathbf{x}_i\beta_i + \mathbf{F}\gamma_i + \mathbf{r}_i \\ \mathbf{r}_i &\sim \mathrm{MVN}\left(0, \sigma_G^2\Phi + \sigma_E^2\mathbf{I}\right).\end{aligned}$$

LMM: Linear Mixed-effects Model

“Mixed”: Fixed and Random effects:

$$\mathbf{y} = \alpha_i + \mathbf{x}_i \beta_i + \mathbf{F}\gamma_i + \mathbf{r}_i$$
$$\mathbf{r}_i \sim \mathrm{MVN}\left(0, \sigma_G^2 \Phi + \sigma_E^2 \mathbf{I}\right).$$

- Pros:
  - Handles any kinship matrix Φ

LMM: Linear Mixed-effects Model

“Mixed”: Fixed and Random effects:

$$\mathbf{y} = \alpha_i + \mathbf{x}_i \beta_i + \mathbf{F}\gamma_i + \mathbf{r}_i$$
$$\mathbf{r}_i \sim \text{MVN}\left(0, \sigma_G^2 \Phi + \sigma_E^2 \mathbf{I}\right).$$

- Pros:
  - Handles any kinship matrix Φ
- Cons:
  - Most computationally intensive
  - Environment is unmodeled

“RGLS”: Reverse Generalized Least Squares

“RGLS”: Reverse Generalized Least Squares

Objective function handles kinship structure (no covariates):

$$O = (\alpha_i' \mathbf{1} + \beta_i' \mathbf{y} - \mathbf{x}_i)^\intercal \, \Phi^{-1} \, (\alpha_i' \mathbf{1} + \beta_i' \mathbf{y} - \mathbf{x}_i) \, .$$

“RGLS”: Reverse Generalized Least Squares

Objective function handles kinship structure (no covariates):

$$O = \left(\alpha_i' \mathbf{1} + \beta_i' \mathbf{y} - \mathbf{x}_i \right)^\intercal \Phi^{-1} \left(\alpha_i' \mathbf{1} + \beta_i' \mathbf{y} - \mathbf{x}_i \right).$$

Effect size estimator (minimizes O) is linear in genotype! (FAST!)

$$\hat{\beta}_i' = \left(\frac{\left(\mathbf{1}^\intercal \Phi^{-1} \mathbf{1} \right) \left(\mathbf{y}^\intercal \Phi^{-1} \right) - \left(\mathbf{1}^\intercal \Phi^{-1} \mathbf{y} \right) \left(\mathbf{1}^\intercal \Phi^{-1} \right)}{\left(\mathbf{1}^\intercal \Phi^{-1} \mathbf{1} \right) \left(\mathbf{y}^\intercal \Phi^{-1} \mathbf{y} \right) - \left(\mathbf{1}^\intercal \Phi^{-1} \mathbf{y} \right)^2} \right) \cdot \mathbf{x}_i.$$

“RGLS”: Reverse Generalized Least Squares

Objective function handles kinship structure (no covariates):

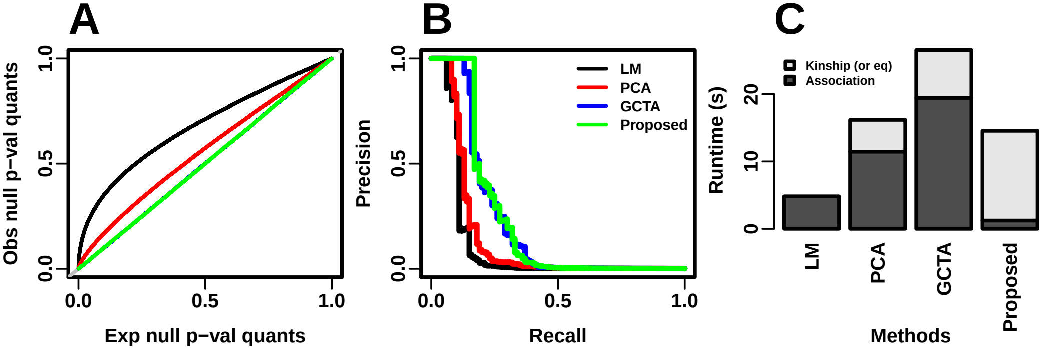
$$O = \left(\alpha_i' \mathbf{1} + \beta_i' \mathbf{y} - \mathbf{x}_i\right)^\intercal \Phi^{-1} \left(\alpha_i' \mathbf{1} + \beta_i' \mathbf{y} - \mathbf{x}_i\right).$$

Effect size estimator (minimizes O) is linear in genotype! (FAST!)

$$\hat{\beta}_i' = \left(\frac{\left(\mathbf{1}^\intercal \Phi^{-1} \mathbf{1}\right) \left(\mathbf{y}^\intercal \Phi^{-1}\right) - \left(\mathbf{1}^\intercal \Phi^{-1} \mathbf{y}\right) \left(\mathbf{1}^\intercal \Phi^{-1}\right)}{\left(\mathbf{1}^\intercal \Phi^{-1} \mathbf{1}\right) \left(\mathbf{y}^\intercal \Phi^{-1} \mathbf{y}\right) - \left(\mathbf{1}^\intercal \Phi^{-1} \mathbf{y}\right)^2} \right) \cdot \mathbf{x}_i.$$

Variance has closed form, yields accurate p-values!

Genetic association on simulated admixture + 20G family



https://github.com/OchoaLab/simtrait

https://github.com/OchoaLab/genio

Conclusions

New genetic association approach:

Conclusions

New genetic association approach:

- Designed for any form of relatedness
  - Works for admixture + family

Conclusions

New genetic association approach:

- Designed for any form of relatedness
  - Works for admixture + family
- Accurate statistics (control false positives)

Conclusions

New genetic association approach:

- Designed for any form of relatedness
  - Works for admixture + family
- Accurate statistics (control false positives)
- Powerful (most true positives, along with LMMs)

Conclusions

New genetic association approach:

- Designed for any form of relatedness
  - Works for admixture + family
- Accurate statistics (control false positives)
- Powerful (most true positives, along with LMMs)
- Faster than competitors

Conclusions

New genetic association approach:

- Designed for any form of relatedness
  - Works for admixture + family
- Accurate statistics (control false positives)
- Powerful (most true positives, along with LMMs)
- Faster than competitors

Next: real data analysis

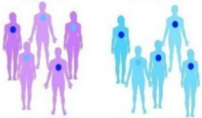
- Duke CARRIAGE family data
- dbGaP datasets

Other projects in the lab

Admixture

Other projects in the lab

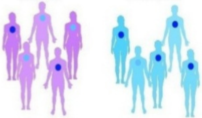
Admixture



Genetic Association Studies

Other projects in the lab

Admixture

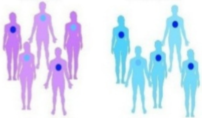


Genetic Association Studies

Heritability

Other projects in the lab

Admixture



Genetic Association Studies

Heritability

Selection

Acknowledgments

Ochoa Lab
Yiqi Yao
Amika Sood

Duke University
Beth Hauser
Yi-Ju Li
Andrew Allen
Amy Goldberg

Princeton University
John D. Storey

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