Population Kinship and Differentiation in Human Studies

Alejandro Ochoa

—

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

StatGen, CBB, B&B — Duke University

2020-02-18 — UPGG seminar

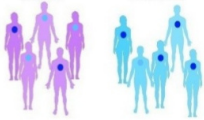
Why study relatedness?

Why study relatedness?

Human genetics is fascinating!

Why study relatedness?

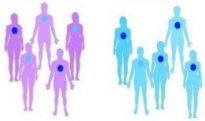
Human genetics is fascinating!



Genetic Association Studies confounded by relatedness

Why study relatedness?

Human genetics is fascinating!

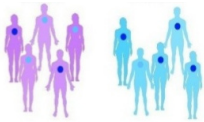


Genetic Association Studies confounded by relatedness

Heritability of complex traits

Why study relatedness?

Human genetics is fascinating!



Genetic Association Studies confounded by relatedness

Heritability of complex traits

Selection scans

Overview

New population kinship and F_{ST} estimates

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

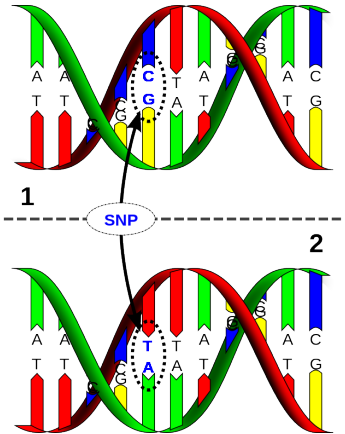
Admixture model

- Hispanics in 1000 Genomes Project
- Infering admixture from a population kinship matrix

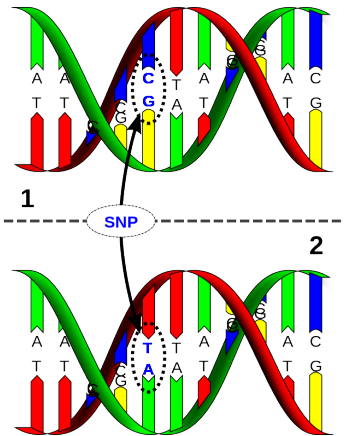
CARRIAGE family study

Inbreeding or deletions in schizophrenia patients?

Single Nucleotide Polymorphism (SNP) data



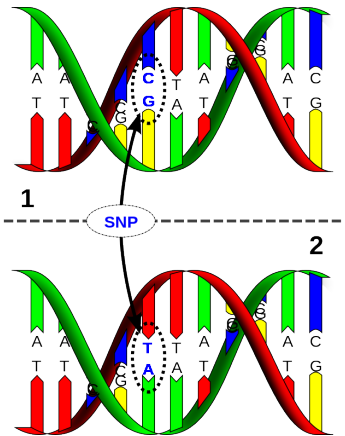
Single Nucleotide Polymorphism (SNP) data



$\Rightarrow$

Genotype	x_{ij}
CC	0
CT	1
TT	2

Single Nucleotide Polymorphism (SNP) data



$\Rightarrow$

Genotype	x_{ij}
CC	0
CT	1
TT	2

$\Rightarrow$

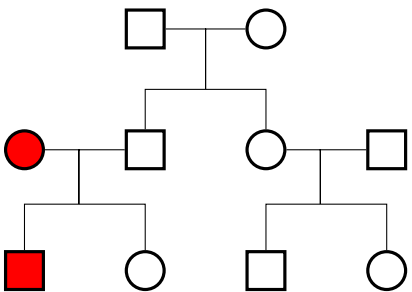
Individuals

Loci

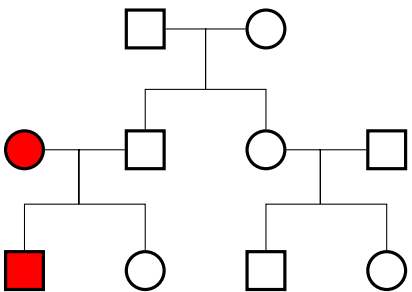
0 2 2 1 1 0 1
0 2 1 0 1
2 ...

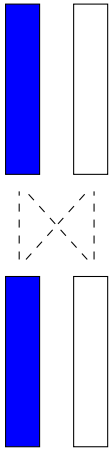
X

The kinship coefficient for parent-child: $\frac{1}{4}$

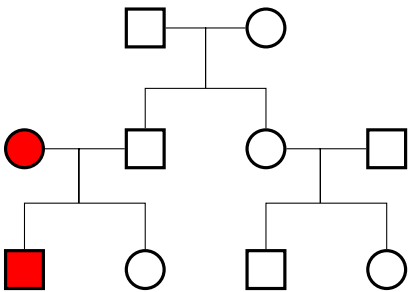


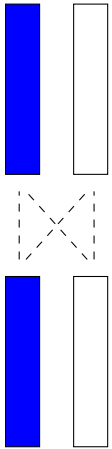
The kinship coefficient for parent-child: $\frac{1}{4}$

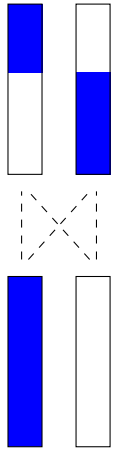




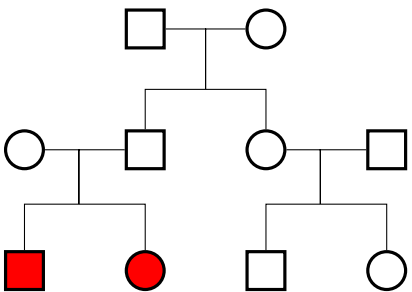
The kinship coefficient for parent-child: $\frac{1}{4}$



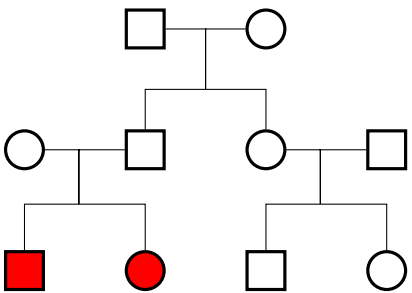


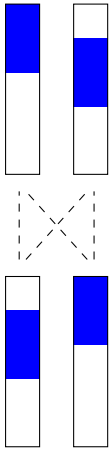


The kinship coefficient for siblings: $\frac{1}{4}$ on average

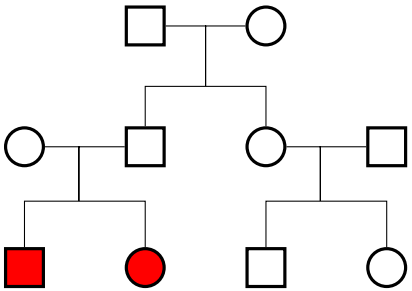


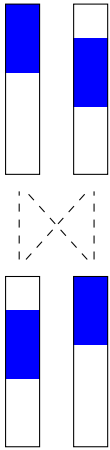
The kinship coefficient for siblings: $\frac{1}{4}$ on average

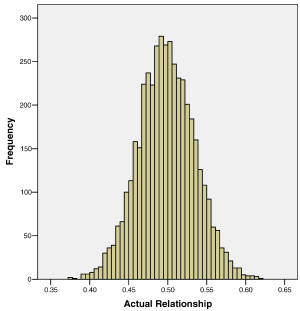




The kinship coefficient for siblings: $\frac{1}{4}$ on average

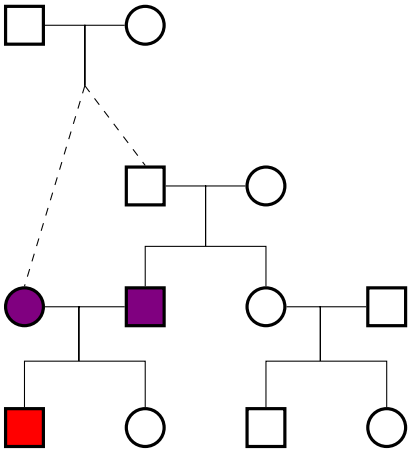




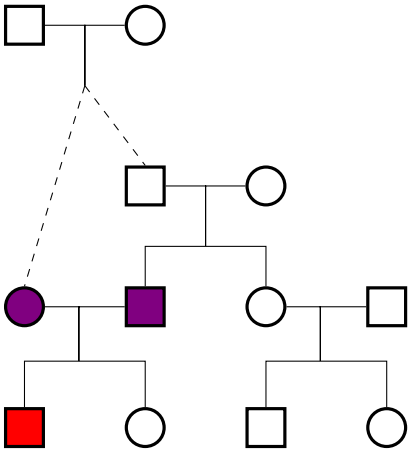


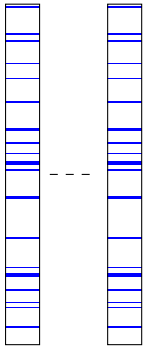
Visscher *et al.* (2006)

The inbreeding coefficient in populations

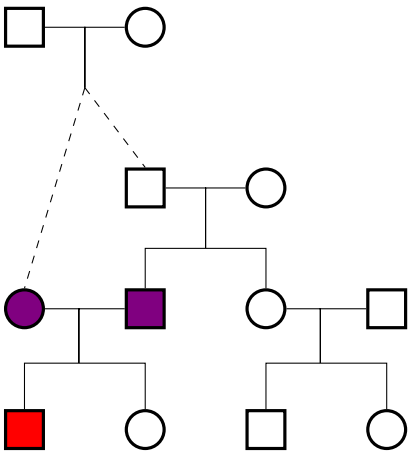


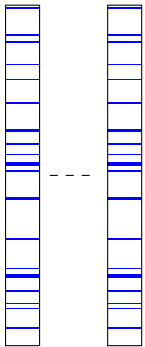
The inbreeding coefficient in populations





The inbreeding coefficient in populations





Measurements relative to a reference pop.:

Inbreeding $= 0$ in the local population

Inbreeding ≥ 0 relative to a distant ancestral population

Better measured using covariance

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

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**:

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

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**:

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

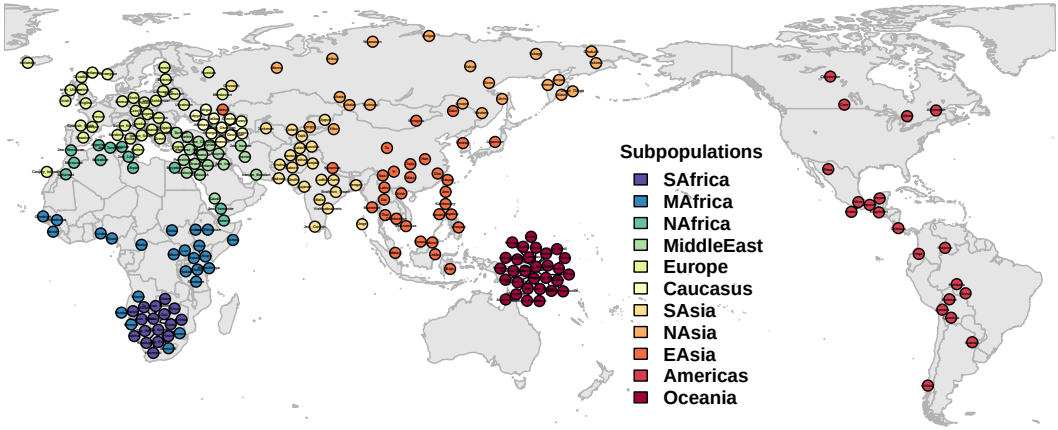
`popkin`: first unbiased kinship estimator! — R package on CRAN

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

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

https://github.com/StoreyLab/popkin

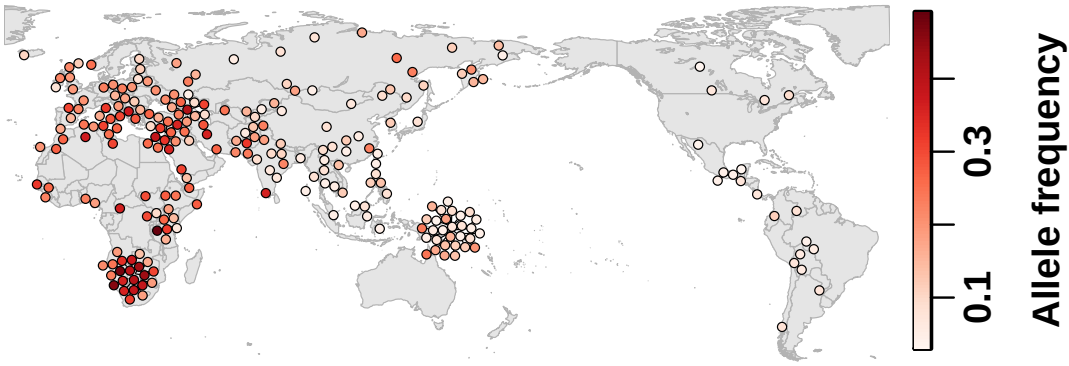
Dataset: Human Origins



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

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

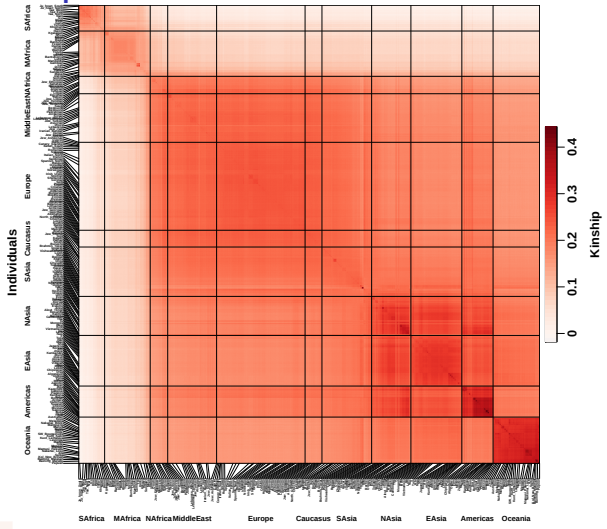
Median-differentiation human locus

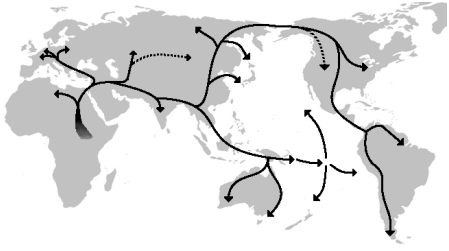


Ochoa and Storey (2019a) doi:10.1101/653279

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

Kinship matrix of world-wide human population



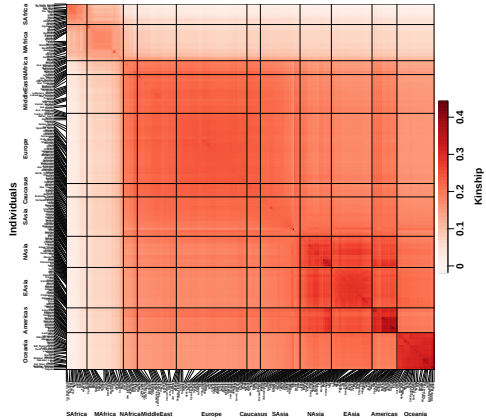


Ochoa and Storey (2019b) doi:10.1101/653279

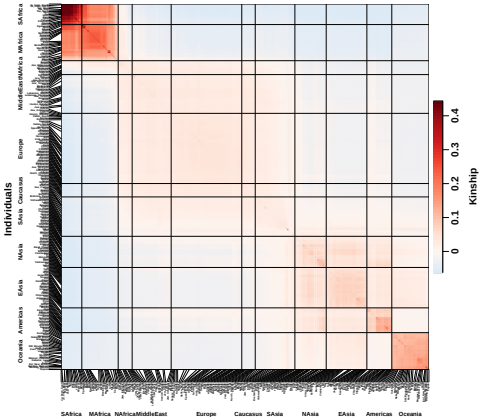
POPKIN https://github.com/StoreyLab/popkin

Standard kinship estimator is severely biased

New



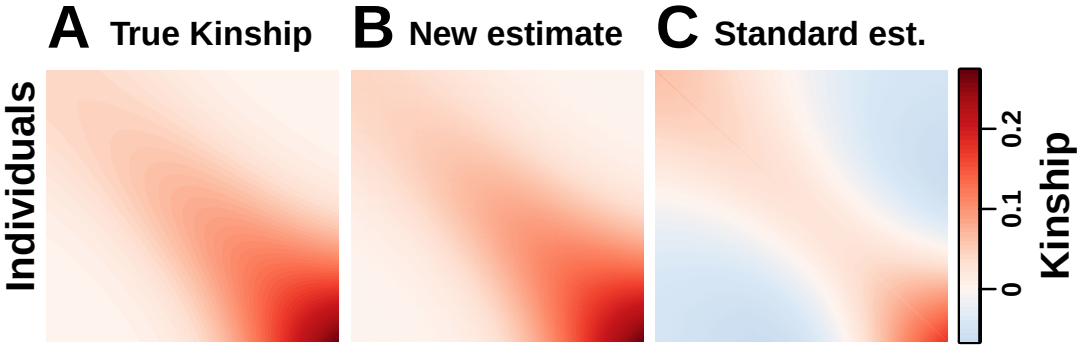
Standard



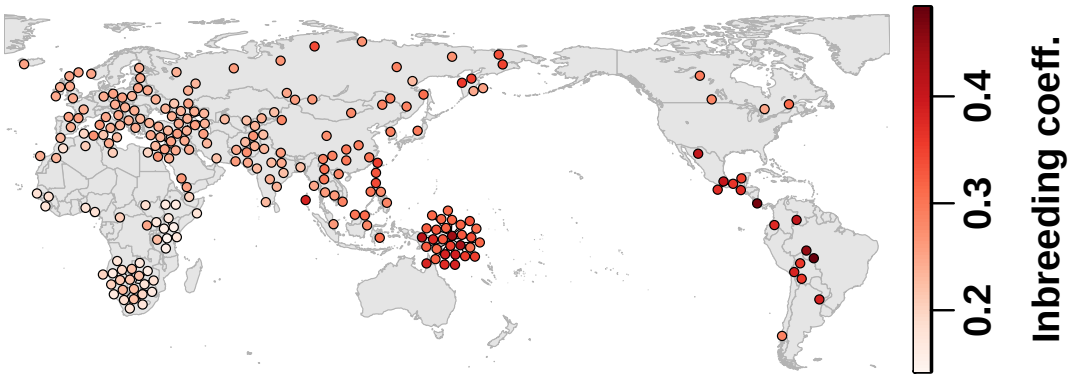
Ochoa and Storey (2019b) doi:10.1101/653279

https://github.com/StoreyLab/popkin

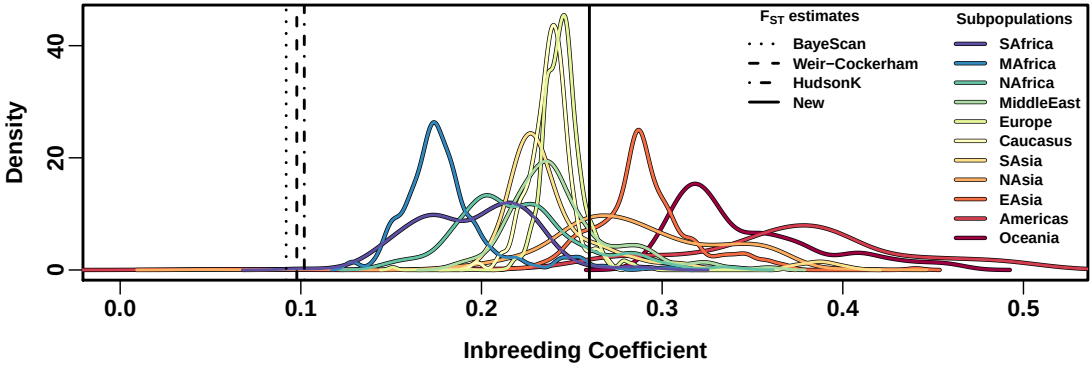
Validation in simulation



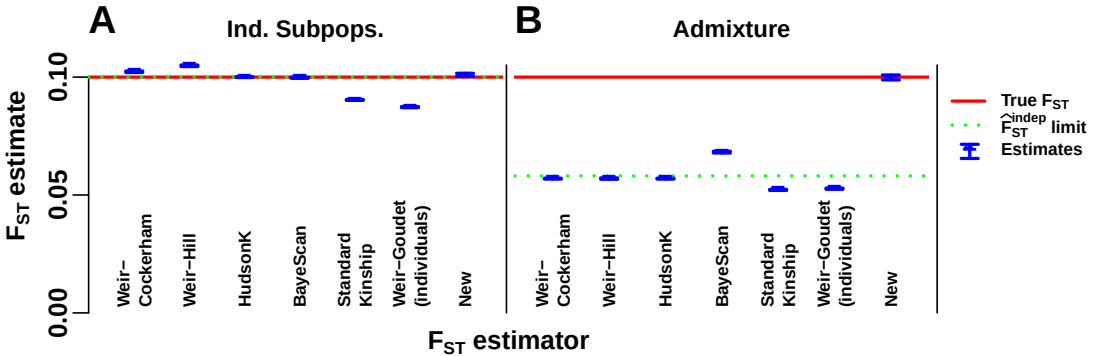
Population-level inbreeding increases with distance from Africa



Differentiation (F_{ST}) previously underestimated



Validation in simulation



Overview

New population kinship and F_{ST} estimates

- Human Origins dataset
- Simulation validations

Admixture model

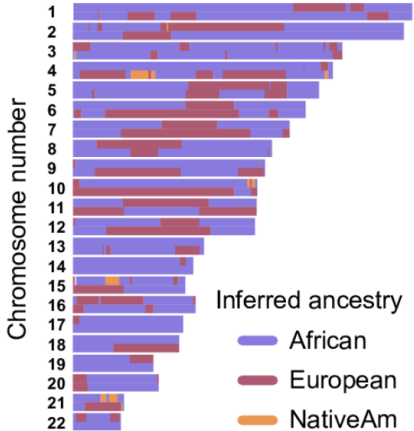
- **Hispanics in 1000 Genomes Project**
- **Infering admixture from a population kinship matrix**

CARRIAGE family study

Inbreeding or deletions in schizophrenia patients?

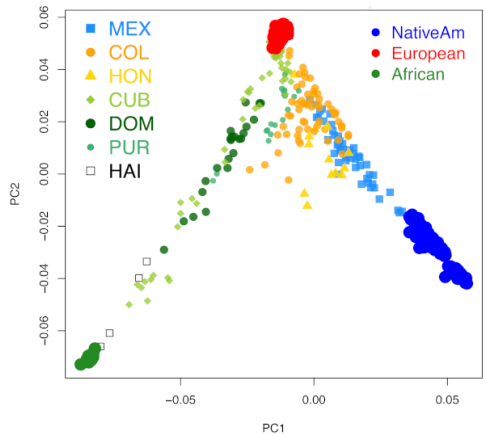
Recently-admixed populations

African-Americans



Baharian *et al.* (2016)

Hispanics



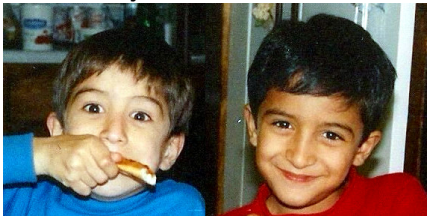
Moreno-Estrada *et al.* (2013)

Admixed siblings from different subpopulations?

Lucy and Maria, UK

Admixed siblings from different subpopulations?

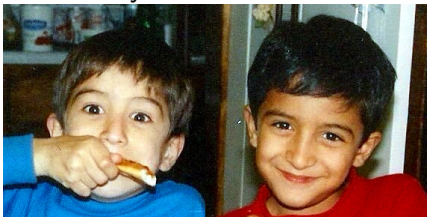
Lucy and Maria, UK



Ochoa brothers, MX

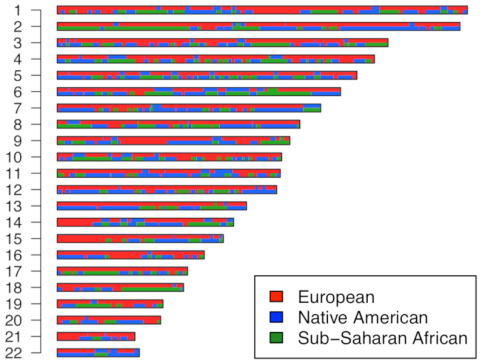
Admixed siblings from different subpopulations?

Lucy and Maria, UK



Ochoa brothers, MX

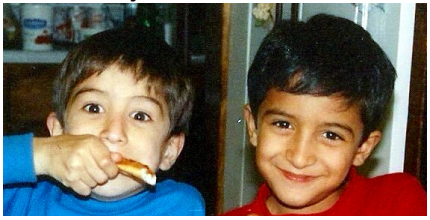
High Admixture LD:



Moreno-Estrada *et al.* (2013)

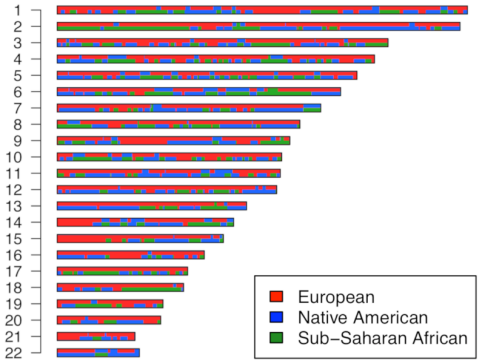
Admixed siblings from different subpopulations?

Lucy and Maria, UK



Ochoa brothers, MX

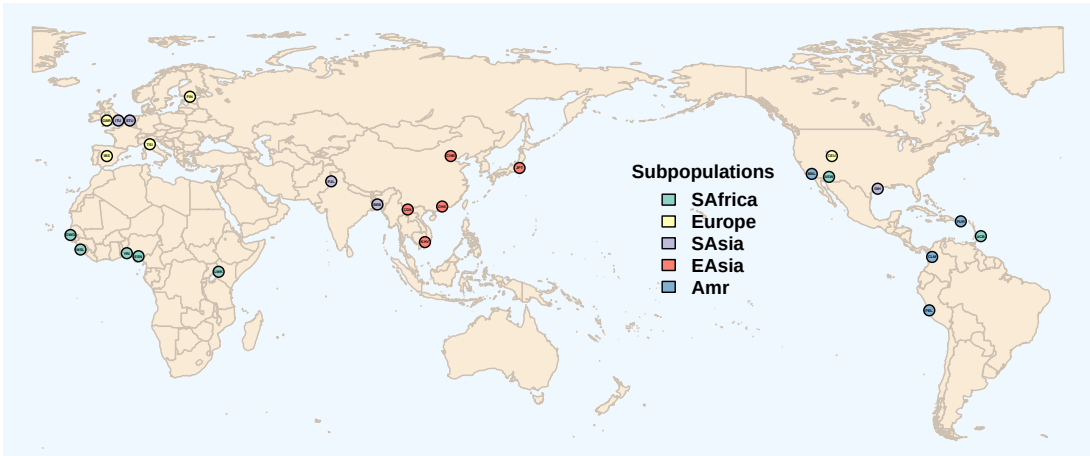
High Admixture LD:



Moreno-Estrada *et al.* (2013)

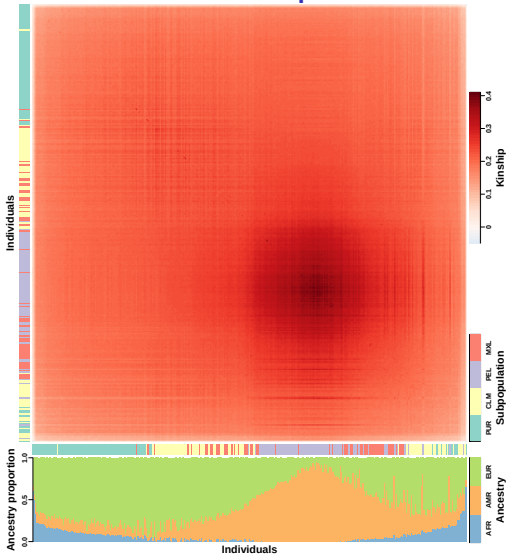
Solution: treat every individual as their own subpopulation!

Dataset: 1000 Genomes Project (2013)



2,504 indivs. from 26 locs. — 20,417,484 loci (asc. in YRI) — WGS, trios, etc.

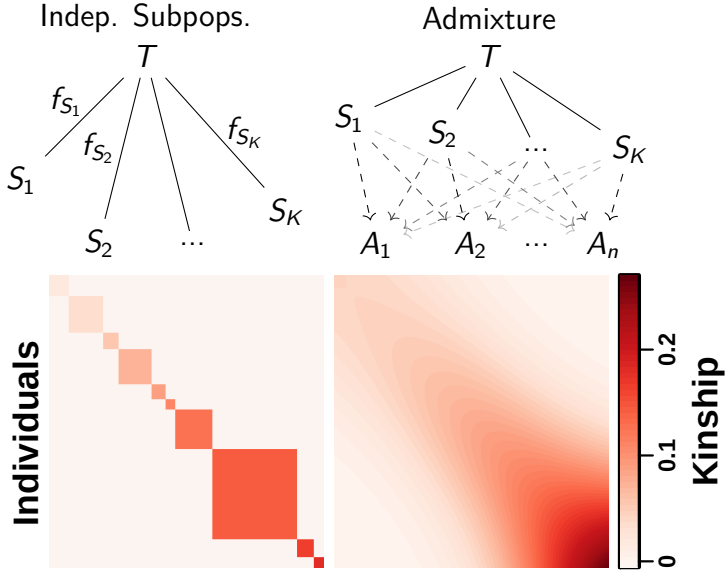
Kinship driven by admixture in Hispanics



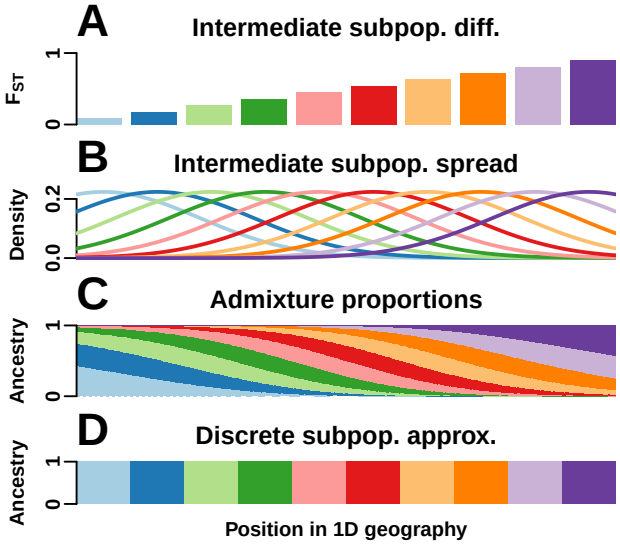
Ochoa and Storey (2019b) doi:10.1101/653279

https://github.com/StoreyLab/popkin

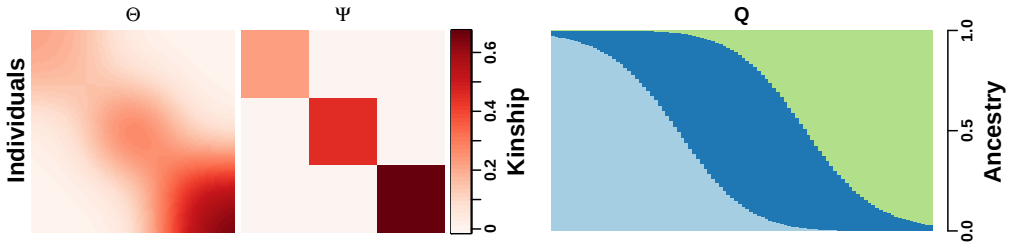
Admixture model



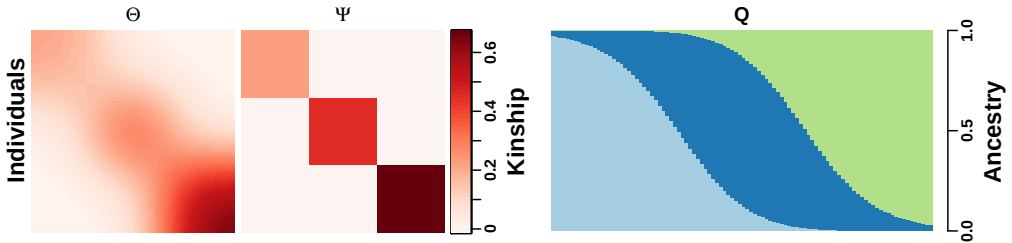
Our admixture simulation



Kinship under the admixture model

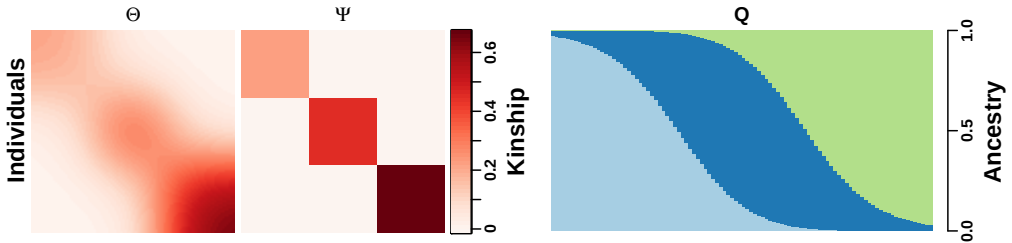


Kinship under the admixture model



$$\Theta = Q\Psi Q^{\intercal}$$

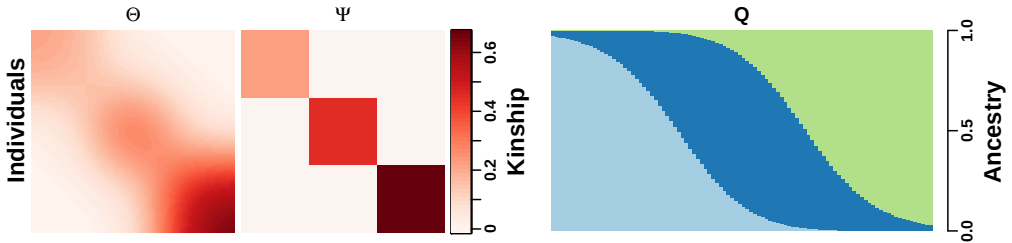
Kinship under the admixture model



$$\Theta = Q\Psi Q^{\intercal}$$

Can we reverse this formula? Estimate admixture proportions from an estimated kinship matrix?

Kinship under the admixture model

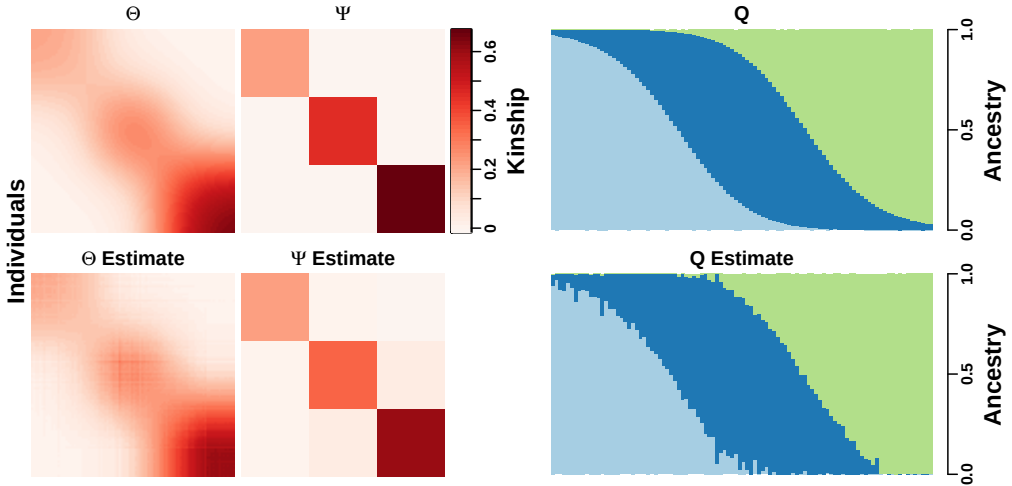


$$\Theta = Q\Psi Q^{\intercal}$$

Can we reverse this formula? Estimate admixture proportions from an estimated kinship matrix?

Why? (1) To understand model, constraints. (2) Could be faster!

Kinship under the admixture model



Yes! Can find least-error fit to kinship, with basic constraints (non-neg, sum to 1)!

Overview

New population kinship and F_{ST} estimates

- Human Origins dataset
- Simulation validations

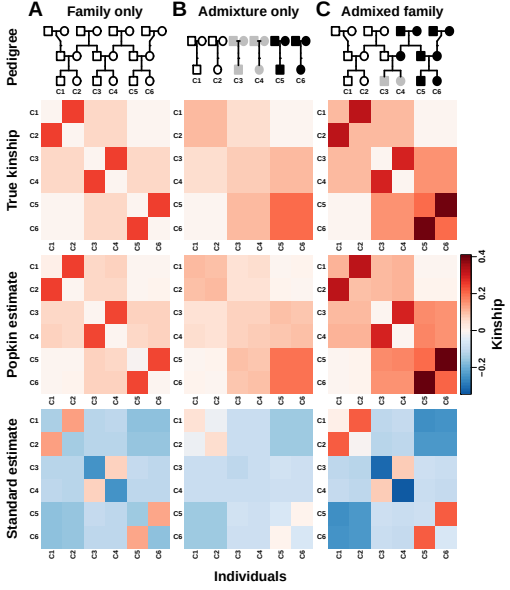
Admixture model

- Hispanics in 1000 Genomes Project
- Infering admixture from a population kinship matrix

CARRIAGE family study

Inbreeding or deletions in schizophrenia patients?

Unified kinship model: ancestry + family structure!

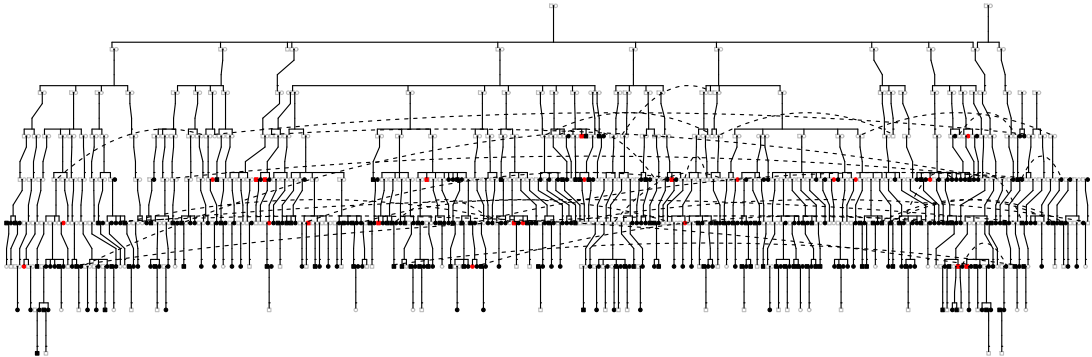


CARRIAGE family: dataset overview

- Local NC family
- African-American, Admixed
- Interested in genealogy, genetics, medicine
- Not ascertained for a particular disease
- Available data:

898	Individuals in pedigree
332	Individuals genotyped
5682	Linkage markers (SNP genotypes)

CARRIAGE family: known pedigree, founders ~1790

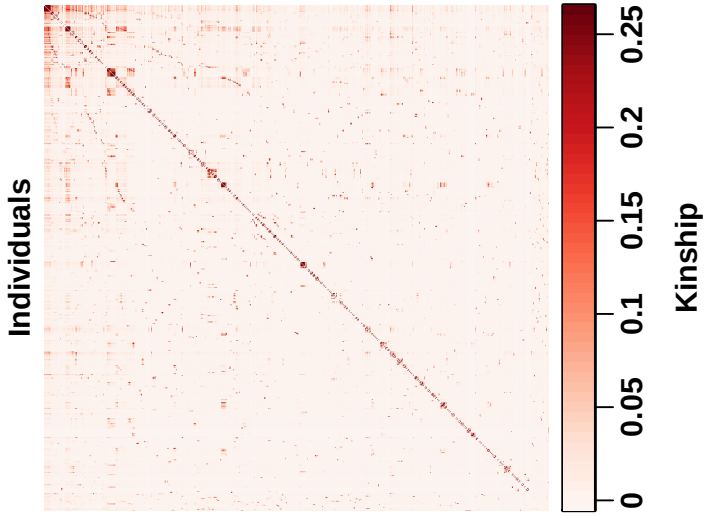


Dashed lines connect copied individuals.

Cardiovascular disease status: unknown, unaffected, affected.

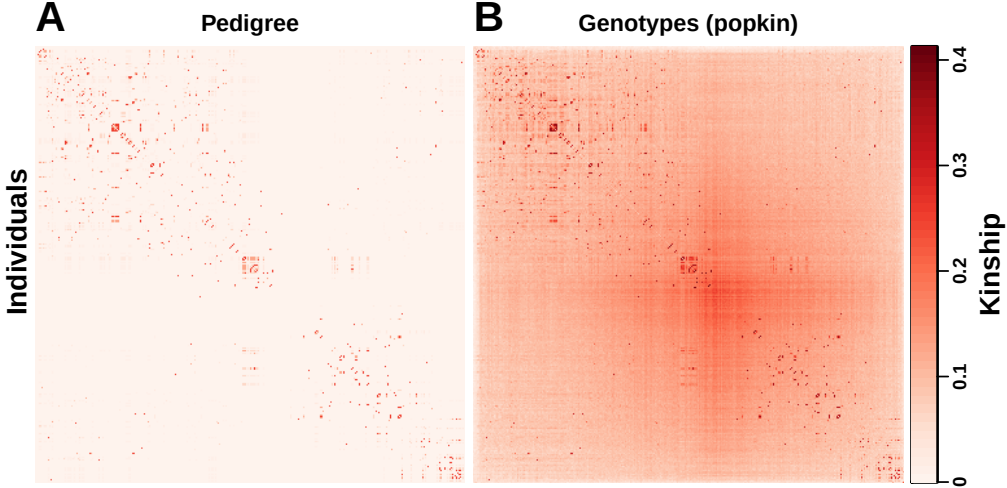
Other phenotypes available.

CARRIAGE family: kinship estimated from pedigree



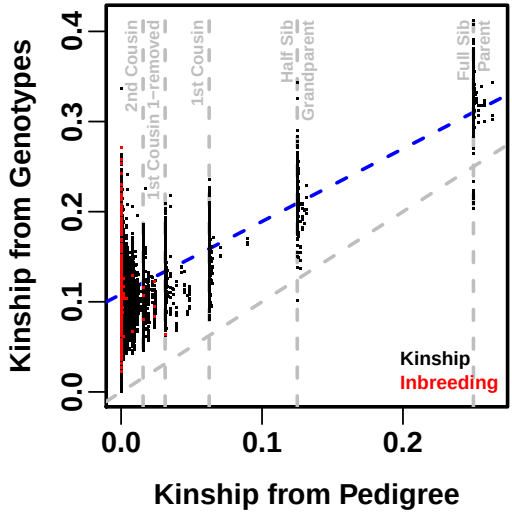
Individuals ordered by pedigree

CARRIAGE family: kinship from pedigree vs genotypes agree!



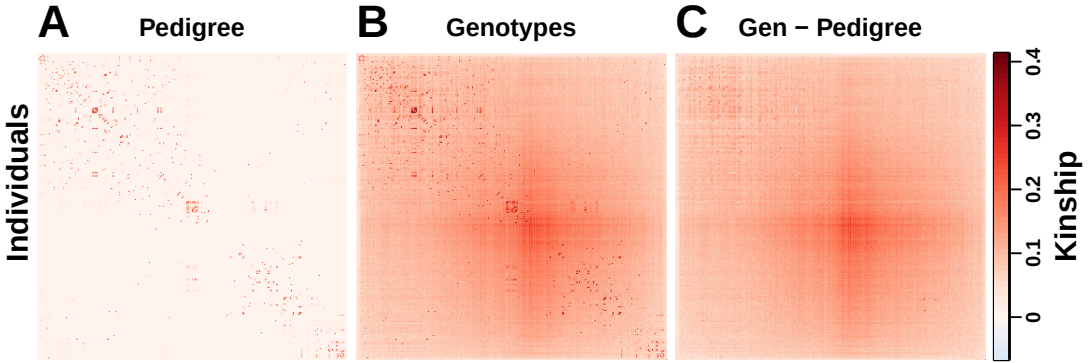
Individuals ordered by seriation

CARRIAGE family: kinship from pedigree vs genotypes agree!



Popkin estimates capture additional kinship due to ancestry

CARRIAGE family: subtracting pedigree reveals ancestry?



Overview

New population kinship and F_{ST} estimates

- Human Origins dataset
- Simulation validations

Admixture model

- Hispanics in 1000 Genomes Project
- Infering admixture from a population kinship matrix

CARRIAGE family study

Inbreeding or deletions in schizophrenia patients?

Otsuka SZ/BD/MDD: dataset overview

- 400 individuals from Aripiprazole drug trials studies of 3 mental disorders:
  - 189: Schizophrenia (SZ)
  - 105: Bipolar Disorder (BD)
  - 106: Major Depressive Disorder (MDD)

Otsuka SZ/BD/MDD: dataset overview

- 400 individuals from Aripiprazole drug trials studies of 3 mental disorders:
  - 189: Schizophrenia (SZ)
  - 105: Bipolar Disorder (BD)
  - 106: Major Depressive Disorder (MDD)
- Demographics:
  - 55% male
  - USA centers. 69% European-American, 26% African-American, 2% Asian, 1% Native-American, $< 1\%$ Pacific Islander and Other.
  - 10% Hispanic

Otsuka SZ/BD/MDD: dataset overview

- 400 individuals from Aripiprazole drug trials studies of 3 mental disorders:
  - 189: Schizophrenia (SZ)
  - 105: Bipolar Disorder (BD)
  - 106: Major Depressive Disorder (MDD)
- Demographics:
  - 55% male
  - USA centers. 69% European-American, 26% African-American, 2% Asian, 1% Native-American, $< 1\%$ Pacific Islander and Other.
  - 10% Hispanic
- 355,542 autosomal loci genotypes (PsychArray)

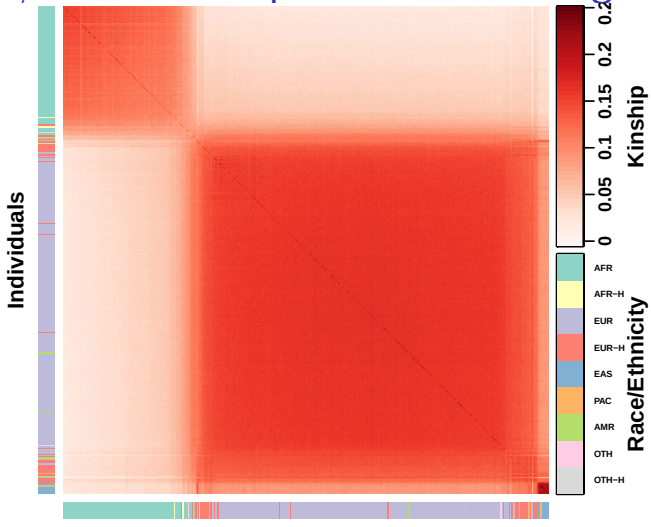
Otsuka SZ/BD/MDD: dataset overview

- 400 individuals from Aripiprazole drug trials studies of 3 mental disorders:
  - 189: Schizophrenia (SZ)
  - 105: Bipolar Disorder (BD)
  - 106: Major Depressive Disorder (MDD)
- Demographics:
  - 55% male
  - USA centers. 69% European-American, 26% African-American, 2% Asian, 1% Native-American, $< 1\%$ Pacific Islander and Other.
  - 10% Hispanic
- 355,542 autosomal loci genotypes (PsychArray)
- Main goal was to model genetics of placebo response (not shown today)

Otsuka SZ/BD/MDD: dataset overview

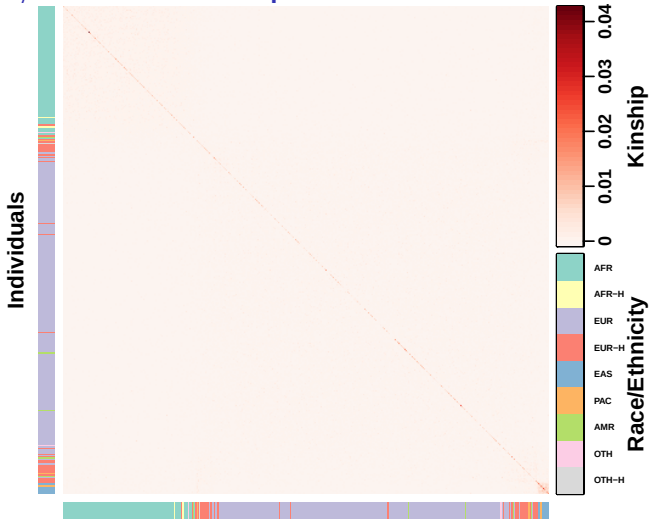
- 400 individuals from Aripiprazole drug trials studies of 3 mental disorders:
  - 189: Schizophrenia (SZ)
  - 105: Bipolar Disorder (BD)
  - 106: Major Depressive Disorder (MDD)
- Demographics:
  - 55% male
  - USA centers. 69% European-American, 26% African-American, 2% Asian, 1% Native-American, $< 1\%$ Pacific Islander and Other.
  - 10% Hispanic
- 355,542 autosomal loci genotypes (PsychArray)
- Main goal was to model genetics of placebo response (not shown today)
- Modeled population structure for GWAS and heritability estimation

Otsuka SZ/BD/MDD: kinship estimated from genotypes



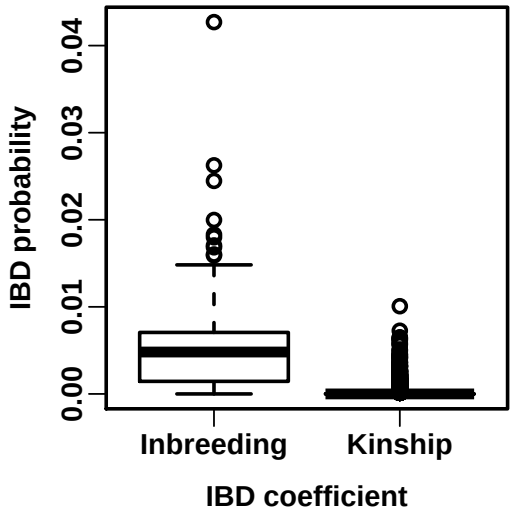
Individuals ordered by seriation

Otsuka SZ/BD/MDD: kinship estimated from IBD blocks



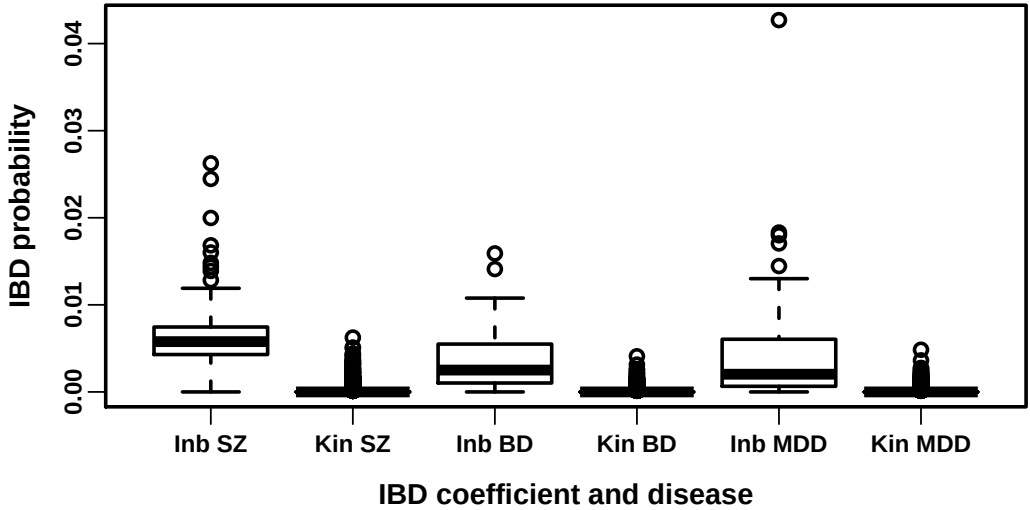
Used Beagle and refined-IBD

Otsuka SZ/BD/MDD: inbreeding greater than kinship?



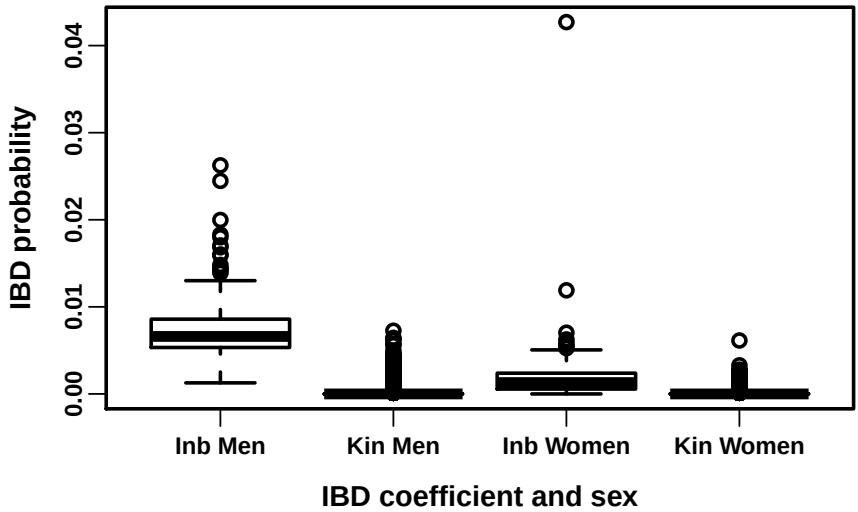
IBD blocks only

Otsuka SZ/BD/MDD: inbreeding vs kinship, by disease



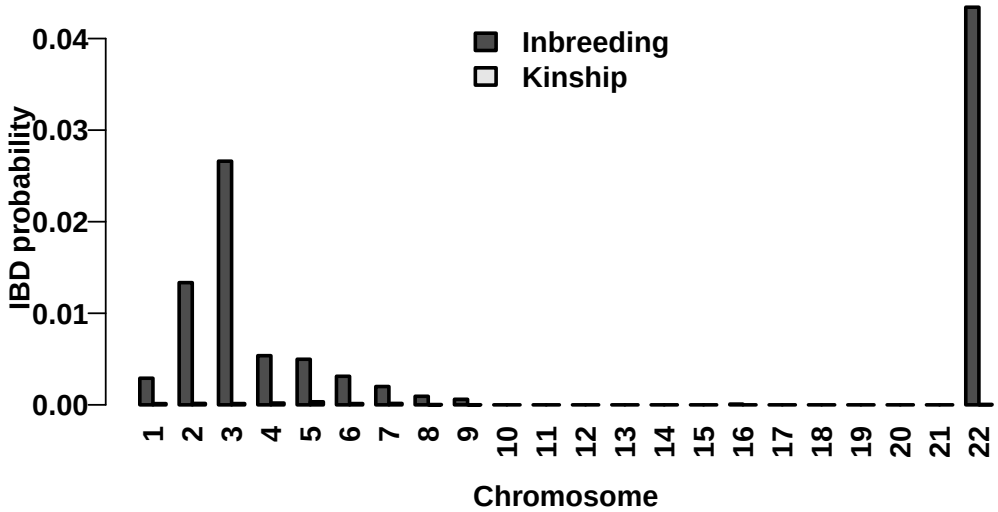
IBD blocks only – Greater in SZ, but present in BD and MDD too!

Otsuka SZ/BD/MDD: inbreeding vs kinship, by sex



IBD blocks only – Greater in men, but present in women too!

Otsuka SZ/BD/MDD: strong chromosomal biases



IBD blocks only – Chr 22 bias suggests 22q11DS cases!

Previous literature: runs of homozygosity (ROH) in SZ

Keller *et al.* (2012): Inbreeding/ROH associated with SZ

Previous literature: runs of homozygosity (ROH) in SZ

Keller *et al.* (2012): Inbreeding/ROH associated with SZ

Johnson *et al.* (2016): Association did not replicate. Confounders?

Previous literature: runs of homozygosity (ROH) in SZ

Keller *et al.* (2012): Inbreeding/ROH associated with SZ

Johnson *et al.* (2016): Association did not replicate. Confounders?

22q11 deletion syndrome, other **chromosomal deletions** associated with SZ (*i.e.* Levinson *et al.*, 2011)

Previous literature: runs of homozygosity (ROH) in SZ

Keller *et al.* (2012): Inbreeding/ROH associated with SZ

Johnson *et al.* (2016): Association did not replicate. Confounders?

22q11 deletion syndrome, other **chromosomal deletions** associated with SZ (*i.e.* Levinson *et al.*, 2011)

Unresolved issues:

- Inbreeding inferred from ROH only
  - Could also be large chromosomal deletions
  - Microarray genotyping does not differentiate ROH causes
- Positional biases ignored

Solution: SZ trio data!

Collaboration with David Goldstein, Columbia University

Solution: SZ trio data!

Collaboration with David Goldstein, Columbia University

- Trio design: genotype affected child and unaffected parents

Solution: SZ trio data!

Collaboration with David Goldstein, Columbia University

- Trio design: genotype affected child and unaffected parents
- Are ROH observed here too?

Solution: SZ trio data!

Collaboration with David Goldstein, Columbia University

- Trio design: genotype affected child and unaffected parents
- Are ROH observed here too?
- If so, can distinguish explanations:
  - Inbreeding: parents must be related, share IBD block in question
  - Otherwise chromosomal deletion

Solution: SZ trio data!

Collaboration with David Goldstein, Columbia University

- Trio design: genotype affected child and unaffected parents
- Are ROH observed here too?
- If so, can distinguish explanations:
  - Inbreeding: parents must be related, share IBD block in question
  - Otherwise chromosomal deletion

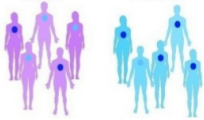
Stay tuned!

Other projects in the lab

Admixture

Other projects in the lab

Admixture



Genetic Association Studies

Other projects in the lab

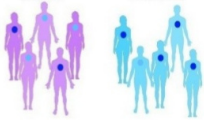
Admixture

Heritability

Genetic Association Studies

Other projects in the lab

Admixture



Genetic Association Studies

Heritability

Selection

Acknowledgments

Ochoa Lab
Amika Sood
Yiqi Yao
Zhuoran Hou

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

Princeton University
John D. Storey

Otsuka Pharmaceutical
Srikanth Gottipati

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